Part B lite (Just the Basics) QA/QC Review Checklist for Aquatic Vital Sign Monitoring Protocols and SOPs

Roy Irwin, NPS, WRD

April 21, 2006 Update

Suggested citation: Irwin, R.J. 2006. Draft Part B lite (Just the Basics) QA/QC Review Checklist for Aquatic Vital Sign Monitoring Protocols and SOPs, National Park Service, Water Resources Division. Fort Collins, Colorado, distributed on Internet only at http://www.nature.nps.gov/water/Vital_Signs_Guidance/Guidance_Documents/PartBLite.pdf

Draft Revisions in Progress. These topics are complex and there will be additional updates and improvements in the future. Please send peer review comments to roy_irwin@nps.gov

Table of Contents (In Word, a control-click on the chapter headings below will take you there):

Disclaimer: Nothing in the discussion below should imply government endorsement (or lack there of) of any specific products. Commonly known products (including books now often used in our field) are mentioned strictly as examples, but there probably others out there that may be superior in various ways now or in the future.

Introduction

The following is condensed from the much longer version of Part B (Irwin, R.J. (2004). Vital Signs Long-Term Aquatic Monitoring Projects: Part B. Planning Process Steps: Issues to consider and then document in a monitoring plan including monitoring protocols and standard operating procedures (SOPs) for Quality Assurance/Quality Control (QA/QC). Water Resources Division, draft available at http://science.nature.nps.gov/im/monitor/protocols/wqPartB.doc.

Some found Part B to be too long, while others complained when it was shortened. Part B is still available as a longer resource and is aimed more at the planning process as whole rather than protocols and SOPs. Part B lite (herein) is an effort to provide networks with a condensed version optimized for use when developing protocol narratives and attached SOPs.

As suggested in generic VS guidance (K.L. Oakley, L.P. Thomas, and S.G. Fancy, 2003. Wildlife Society Bullet 31(4), reprint at http://science.nature.nps.gov/im/monitor/protocols/ProtocolGuidelines.doc), all sampling protocols will include three basic sections:

A. Protocol Narrative
B. Protocol Standard Operating Procedures (SOPs), and
C. Protocol Supplementary Materials

An important item on any protocol review is whether or not the protocol follows the organization above, is complete, and has a table of contents that helps one determine where things are. We recommend that a QA/QC SOP be included that covers most of the topics covered herein. For those networks who want to follow state and EPA conventions, the QA/QC SOP could also be called a quality assurance project plan (QAPP) SOP.

Many of the QA topics covered in the first sections herein are touched on briefly in the protocol narrative, often with more detail in the QA/QC SOP. The QC topics (from comparability on down herein) can be covered primarily in the QA/QC SOP. Many of these topics are interrelated and the different pieces need to make sense when considered as a whole. Therefore, we recommend liberal use of "point-to" links to help readers understand the big picture and where the important pieces are.

Either the protocol narrative or a separate SOP should include a discussion of who will do the monitoring and who will train them and how often (recurrent training and is Quality Assurance/QA basic). Is there a SOP that clearly defines protocol variables and how to measure them?

The following text summarizes the basics of what has to be in water quality and other aquatic protocol SOPs to meet checklist ("Checklist for Review of Vital Signs Monitoring Plans," on the Internet at http://science.nature.nps.gov/im/monitor/docs/MonitoringPlanChecklist.doc, hereafter referred to as "the checklist") requirements.

This summary can also be used for the basics that should be included in Phase 1, 2, and 3 monitoring plan chapters. In most cases, the planning process is iterative, with

very general statements in the plan chapters becoming more detailed in the subsequent protocol narrative, and then even more detailed in SOPs.

Among the basics that need to be covered in the narrative and SOPs are (adapted from USGS wildlife monitoring website at http://testweb-pwrc.er.usgs.gov/monmanual/):

WHAT are you going to measure?
WHERE are you going to put your sampling points?
HOW are you going to measure it?
WHEN (and how frequently) are you going to measure it?

One of the main originators of the survey design and response design concepts popular in EMAP and other survey design disciplines helpfully clarified the terminology distinctions as follows (Scott Urquhart, Department of Statistics, CSU, Personal Communication, 2005):

What: Sampled Population and/or Target Population
Where: Monitoring, Survey, or Sampling Design
How (and Who) -- The Response Design
When (and how often) -- The Temporal Design

In modern scientific thinking (as well as modern environmental monitoring planning), quality assurance is not just a last minute task one does at the end of planning, but includes the entire planning process, including carefully thinking through the questions that need to be answered after summarizing what is already known, making sure the data collected are relevant, representative, comparable, and of adequate quality and quantity, and making sure the study design is defendable and "makes sense." See Part B (the longer version) for: 1) more detailed discussions of the most of the topics herein, 2) additional detail on the differences between quality assurance and quality control, 3) much more detailed discussions of the entire planning process and how all the pieces fit together.

.

QUALITY ASSURANCE:

Summary of Information from Past Data

QA checklist question: For water quality monitoring, has information content of available past aquatic data (for each waterbody being considered for monitoring) been adequately summarized in terms of hints of trends or other important issues of concern? The word hint is used since old data is seldom perfectly definitive, complete, or perfectly comparable between agencies or time periods.

The emphasis should not only on what groups have been monitoring where and when, but also on "what does the data collected mean?" Again, what is the information content of the past data regarding hints of trends or issues of concern? Although this may have been briefly mentioned in chapter 1 of the central monitoring plan, typically more detail should be provided in protocol narratives.

A table listing 303d waters should be included in the protocol narrative, along with a note that the most recent WRD Designated Use and Impairments database (at http://www1.nrintra.nps.gov/wrd/dui/) has been consulted and that any differences with the vital sign network versions of the 303d lists have been logically reconciled. When possible, there should be more spatial detail (impaired from where to where?) in protocol narratives compared to related discussions in the background section in chapter 1 of the central monitoring plan.

Document Objectives and Questions

It is easiest to plan monitoring if the general objectives in the central monitoring plan are rephrased into more detailed questions in the protocol narrative. A QA basic is that if the questions are sufficiently detailed, it monitoring can be planned in such a way that questions can be answered with the data collected. As monitoring protocols and SOPs are revised, it is important that the final more-detailed questions continue to make sense in comparison with summary discussions for representativeness and named target populations (or sampled populations) about which inferences will be made.

In the same general section where questions are being detailed, each protocol narrative should address the following checklist question: "Does the protocol narrative identify specific measurable objectives such as thresholds or trigger points for management actions?" When possible the details of any such thresholds or trigger points should be fully explained in the protocol narrative.

As was the case for questions, thresholds and trigger points should be discussed briefly in the protocol narrative and addressed in more detail in SOPs (such as the QC or Data Analysis SOP).

For example, are the thresholds of concern water quality standards? If the threshold or comparison benchmark is a water quality standard that already has some safety margin built in, managers may still want to know when a standard is being closely approached. Therefore, the magnitude of the change that needs to be detected in trend analysis detect should typically be smaller than the entire distance between current condition and the standard.

Is the threshold to be used a resource-collapse threshold value with no safety margin? If so, an even bigger safety margin would usually need to be factored into decisions about how big of a change needs to be detected. What units will the safety margin use?

Being able to detect an effect size of concern typically depends on variability of the data, sample size, alpha, and beta. These plus safety factors are input variables used to determine sample sizes and data completeness (completeness should be covered in the QA/QC SOP).

Safety factors and threshold should be covered at least briefly in the protocol narrative (see data completeness discussion below in the subsection on developing a safety margin between existing conditions and threshold magnitudes). If the network places details on these issues in the Data Analyses SOP or the QA/QC SOP, the network should also place a "point-to" marker in each protocol narrative so that readers can more easily find the more detailed discussions.

The protocol narrative should also summarize which questions and/or sites were selected to ensure monitoring of a 303d impaired water body or a very pristine water body that the park wants to keep that way. WRD has suggested that at roughly 2/3 of the sites should be in one of those two categories (see Part A guidance). What monitoring will be done to help answer GPRA reporting goals?

Document Vital Signs and Measures and How They Were Chosen

The protocol narrative should have a brief recap (or point to where the information may be found) on what will be measured and how vital signs and measures were selected. Was a set of neutral selection criteria used, such as those listed in Part B and in Kurtz et al. (J. C. Kurtz, Jackson, L. E., and W. S. Fisher. 2001. Strategies for evaluating indicators based on guidelines from the Environmental Protection Agency's Office of Research and Development, Ecological Indicators 1:49–60, http://science.nature.nps.gov/im/monitor/docs/EvalEcolIndic.pdf)? Is so, what were the criteria?

A QA basic is that measures chosen should be helpful in answering the questions. There should be a brief explanation in each protocol narrative of how the measures (typically level 3 Vital Signs) selected relate to both 1) values to be protected, and 2) desired conditions/ecological relevance. Which vital signs or measures were picked due to regulatory water quality impairment (303d lists, GPRA, etc., see Part A and B for more detail)?

Selected measures should ideally be simple and explained in plain language in the protocol narrative (see Australian example of simple explanations at http://apostle.environment.wa.gov.au/idelve/srwqa/methodology.htm).

As explained in more detail below in the section on completeness and statistics, other things being equal, measures picked should ideally not have:

1. Very poor measurement uncertainty (poor measurement precision and/or unacceptable measurement bias), or
2. Extremely high true variability at relatively un-impacted sites.

For programs with limited monitoring budgets (= small sample sizes), excesses in either of these (or both) can prevent the detection of a true change of a magnitude of concern, or the detection of a standards exceedance.

The neutral criteria used in picking vital signs and measures (such as those listed just above) should be summarized in the protocol narrative.

As explained in more detail in Part B and the National Park Service (NPS) freshwater and marine white papers (http://www.nature.nps.gov/water/VitalSigns_index/VitalSignsdocuments.cfm), several core parameters are required any time aquatic sampling is done:

> For freshwater, required parameters include specific conductance (differs from conductivity by being temperature corrected), dissolved oxygen, pH, and water temperature. In addition, photographic documentation of the collection site (a minimum record of one digital site photo) is recommended, along with at least a

STORET-terminology-consistent qualitative assessment of flow "severity" (choice list from http://www.nature.nps.gov/water/infoanddata/index.cfm):

Choices	Description
DRY	No visible water in stream (typical of dry period for an ephemeral/intermittent stream).
NO FLOW	Discrete pools of water with no apparent connecting flow (at surface).
LOW	Base flow for a stream or flow within roughly 10% to 20% of base flow condition.
NORMAL	When stream flow is considered normal (greatest time that stream is characterized by this in terms of flow quantity, level, or general range of flow during a falling or rising hydroperiod, but above base flow).
ABOVE NORMAL	Bank full flow or approaching bank full (generally within upper 20% of bank full flow condition).
FLOOD	Flow extends outside normal bank full condition or spreads across floodplain.

Except for "low flow", similar terminology could be used for lakes, ponds, reservoirs, or wetland water levels, though the terminology is not now standardized in STORET. If enough networks agree on terminology, we could suggest new terminology for STORET. For example, the network might choose a rating such as the following expressed a % of bank full:

- Low - (<25% of bank full)
- Intermediate (or normal?) - (25% ≤ Q ≤ 75% bank full)
- High (or above normal?) - (> 75% bank full)
- Flood Stage (overbank) - (> 100% bank full)

If a more complex lake, pond, or wetland rating system is used, something relative simple (like the above) might still be used in addition to the more complex terminology. For example, the EPA lakes habitat protocol at http://www.epa.gov/emap/html/pubs/docs/groupdocs/surfwatr/field/lake_hab.pdf mentions lake levels in three places:

1) Estimate the vertical and horizontal distances between the present lake level and the high water line.

2) The riparian habitat characterization includes riparian vegetation cover, shoreline substrate, bank type and evidence of lake level changes,

3) (Section 5.2.2.1.3): Bank type and evidence of lake level changes--Choose the bank angle description that best reflects the current shoreline that is dominant within your field of vision and 1 m into the riparian plot: V = Near vertical/undercut (>75 degrees, S = Steep; >30 to 75 degrees, hard to walk up bank; or G = Gradual, 0 to 30 degrees, easy to walk up). Estimate the vertical difference between the present level and the high water line; similarly, estimate the horizontal distance up the bank between current lake level and evidence of higher level.

If the waterbody is dry, other water column parameters like pH and conductivity cannot be taken, but recording the fact that the habitat is dry may be important to tracking changes in frequencies of flow or water level conditions. Changing stream flows was singled out as an especially important national freshwater ecological indicator in the Heinz Report (The Heinz Center. 20003. State of the Nation's Ecosystems: Annual Update 2003 (http://www.heinzctr.org/ecosystems/fr_water/indicators.shtml).

If the Cowardin et al (1979) wetland classifications for hydrologic regime (saturated, temporarily flooded, seasonally flooded, semi-permanently flooded, permanently flooded, etc.) are used to describe the type of wetland, that should be done in addition to rather instead of the instantaneous water level qualitative terms such as those above. In other words, it is still useful to know how full the wetland was when making water quality of aquatic biology measurements in a wetland.

For marine or estuarine monitoring, required parameters include ionic strength expressed as conductivity and as salinity, pH, dissolved oxygen, and water temperature. In addition, the following are required:

- Location standard coordinates [GPS on collection sites and also consult the Universal Transverse Mercator (UTM) grid; on USGS quad];
- Local time (indicating standard or daylight-saving time);
- Water depth and sample depth;
- Tidal stage (e.g. high, low, or mid-tide) and direction (ebb, flood or slack water),
- Estimated Wave Height.
- Flushing time
- Tidal range
- Habitat description

Include Detailed SOPS for All Field and Lab Methods

A QA basic is that methods should be explained in detail. Exactly what will be done in the field and lab? Reproducibility and transparency are not only QA basics but also sound science basics. The amount of detail in the SOPs should be sufficient so that someone outside the NPS could duplicate the methods exactly.

As required by Oakley et al. 2003 (op.cit.), various NPS WRD guidance documents, and modern QA/QC conventions, all protocols should include method-detail

SOP(s). For convenience, the method SOPs may be broken down into two groups: A field method SOP and a lab method SOP.

Together, the SOPs should fully explain the planned "response design," the process of obtaining a response at a site, once sites have been chosen (see EMAP at http://www.epa.gov/nheerl/arm/designpages/response_design.htm).

Will USGS National Water-Quality Assessment Program (NAWQA) or state or EMAP protocols and SOPs be used? What detailed field and lab protocols will be used to get a response at the site? The response design incorporates numerous decisions about how to measure the attribute of interest accurately (Larsen, D. P., T. K. Kincaid, S. E. Jacobs and N. S. Urquhart (2001). Designs for evaluating local and regional scale trends. Bioscience 51:1069-1078 and EPA EMAP definitions at http://www.epa.gov/nheerl/arm/terms.htm).

Some agencies (EMAP and NAWQA) have utilized response designs that call for only a single site visit each year, usually sampled during a narrow index time period. This is typical for monitoring designs conducted over large regional areas. However, one can choose sites with a probabilistic design and then decide to sample the same sites (or with in the same general area or reach) more frequently. Details in a response design are driven by the objectives and questions to be answered by monitoring.

An example of a detailed protocol for a response design for wadeable streams is explained in the EMAP protocols (Lazorchak, J.M., Klemm, D.J., and D.V. Peck (editors). 1998.Environmental Monitoring and Assessment Program-Surface Waters: Field operations and methods for measuring the ecological condition of wadeable streams. EPA/620/R-94/004F. U.S. Environmental Protection Agency, Washington, D.C.). For those in the Western US a resource (albeit still under revision and thus not officially citable) is also available on web (Western Pilot Study: Field Operations Manual for Wadeable Streams, at http://www.epa.gov/emap/html/pubs/docs/groupdocs/surfwatr/field/ws_chap.html).

A diagram of typical EMAP placement of sites along a river (showing alternation of left, middle and right collecting spots) is at http://www.epa.gov/nheerl/arm/designpages/streams/emapstrmresp.htm. Several Networks either have water or aquatic protocols and SOPs in the works based (at least partly) on the Western Pilot model or are considering using it as a model (NCPN, ROMN, NGPN, GRYN).

Does the protocol narrative provide evidence that the monitoring network fully considered using existing protocols or SOPs used by state(s) or large regional monitoring programs (such as USGS, EMAP, or NOAA)? Many existing protocols and especially SOPs can often be adopted and be used as is. Using existing protocols and SOPs to the extent possible is a good idea as it helps with regional data comparability.

In picking methods, a good first step is to scan methods and SOPs in the National Environmental Methods Index (NEMI, www.nemi.gov) to get a relatively quick idea about which ones can achieve true method detection limit (MDLs) lower than all water quality standards or other comparison benchmarks or thresholds of concern. When possible, use methods that have acceptable MDL detection limits rather than RNGE—(range-defined) detection limits.

When possible, choose methods and labs where the semi-quantitative (MDL) detection limit that can be achieved is lower than the lowest water quality standard or

other benchmark. Better when possible; choose methods and labs allowing the MDLs to be 1.6 to 2 times lower than comparison benchmarks. Best when possible, the MDL should be more than 3.18 times lower than the water quality standard or other comparison benchmark. The quantitative limit (ML) detection limit is 3.18 times the MDL, so this is the ideal scenario where both the MDL and ML would be below all data comparison benchmarks, standards, or thresholds.

Getting the absolutely lowest detection limits (better and best examples above) is more important in some situations than in others. If monitoring networks anticipate that many of the their measurements will involve very low level signals (low concentrations near the MDL detection limits), it is worth going to some trouble to find methods and labs that can achieve MDLs that are below the anticipated levels and comparison benchmarks. This is especially true for chemicals that can be a concern when present even in very low amounts and for nutrients in very pristine sites where nutrients are very low. However, for some analytes, methods or labs that can achieve detection limits that low cannot be found. In other cases, all measurements are likely to be well up in the quantitative range. For example in farming or urban areas, nitrate levels in surface water quality are never likely to fall to levels near low-level detection limits. In this case, the lowest possible low-level detection limits may not be needed and the network could consider using methods and labs with higher detection limits if doing so reduced costs.

While one is quickly screening methods in NEMI to see if the method can achieve acceptable detection limits, it would be time-efficient to also check to see if the listed method performance for precision and bias (% recovery) are acceptable for project purposes. If acceptable detection limits, precision performance, and bias performance capabilities are not listed in NEMI or in the method itself, it is reasonable to consider other methods already having acceptable performance documented.

Lab SOPs should detail exactly how everything is done in the lab. If a standard method from a state, USGS, or EPA is used, it should be written out or attached in its entirety and electronic copies should archived the database so that users can find out exactly what was done years from now. If no electronic versions exist, hardcopies should be archived and "point-to" notes in the database should give the location of storage. Method and SOP documentation should include measurement quality objectives (MQOs) for MDLs, Precision, Systematic Error/bias/% recovery (still wrongly called accuracy in some methods), and blank control. Many of the EPA methods are in NEMI and can be copied electronically.

If the agency (EPA, USGS, etc.) changes the method, will the NPS also change in the same way? The SOPs should be detailed enough to allow third parties to reproduce the methods and to allow determinations of data comparability (see further discussions below).

Examples of details to be included in protocol narratives or SOPs rather than in central monitoring plans:

> More details on sampling locations and method specifics.
>
> For example, in chapter 4 of the central monitoring plan a table might say that chlorophyll *a* was the parameter to be monitored. SOPs with each protocol should document the rest of the details. For example, a field methods SOP might clarify

that chlorophyll *a* is to be monitored using field water collection procedures of the USGS field manual. A lab methods SOP might then further clarify that in the USGS national NWQL lab in Denver would do the work using USGS Schedule 1637 method. Alternatively, the lab SOP might specify another method (such as EPA method 445.0 or APHA method 10200H-4), was to be used. The entire method used should be copied and included as appendix to the appropriate SOP.

If flow or water level is to recorded, will it be qualitative or quantitative?

What field instrumentation will be needed?

What pre and post season activities are required?

The field methods SOP should detail how will samples be collected and preserved, what containers will be used, and what maximum holding times were used. Unless otherwise justified, use holding time guidance in 40 CFR Part 136 to 136.3 and appendices.

The QA/QC SOP should detail what will be done with data from samples that exceeded holding time requirements. Will such data be rejected or flagged? Data rejection and re-sampling so that newer replacement samples meet the holding times is usually the better option but flagging may be better than simply using the data as though it was high quality data. If flagging is chosen, it should be justified and flagging should be STORET-compatible as follows:

> In cases where the sample exceeded the recommended holding time, and a decision was made to keep the data, the network can enter the data in STORET with a STORET remark code of "EHT", designating the condition that the "Sample or extract held beyond acceptable holding time." The data can also be entered the same way in NPSTORET (http://www.nature.nps.gov/water/infoanddata/index.cfm#NPSTORET). In NPSTORET, four fields over to the right is the Lab Remarks field where one can select "EHT" and/or other remarks/data qualifiers.

Attach a QA/QC SOP to Each Aquatic Protocol

The checklist states that "For water quality monitoring, there should be a quality control SOP associated with each protocol that adequately documents quality control (QC) objectives for measurement sensitivity (detection limits), measurement precision, measurement systematic error (bias as percent recovery), data completeness (including adequacy of planned sample sizes and statistical power), and (if applicable for lab measurements only) blank control" (http://science.nature.nps.gov/im/monitor/docs/MonitoringPlanChecklist.doc). Since the checklist text is so brief, more information on what the WRD will be looking for in a QA/QC SOP is provided herein.

Again, for aquatic projects, QA/QC details should ordinarily be included in a separate QA/QC SOP attached to each protocol. It would also be optimal to include such a SOP for terrestrial or general biology ecology protocols, since data quality is increasingly considered important and valued for those disciplines as well.

The QA/QC or QAPP SOP should include sections explaining how each the following will be controlled and documented for individual parameters or suites of parameters:

1) Representativeness, Target Population, And Completeness
2) Data Comparability (Internal/NPS And External/Other Regional Data)
3) Measurement Sensitivity (Usually MDL And ML) Detection Limits
4) Measurement Precision As Reproducibility And/Or Repeatability
5) Measurement Systematic Error/Bias/Percent Recovery (Still Wrongly Called Accuracy By Some)
6) Blank Control Bias (Usually For Chemical Lab Work Only)

If additional detail on any of the topics listed above is located anywhere other than in the QA/QC SOP, a summary of what will be done to control each of the issues listed above should be included in the QA/QC SOP. Point-to hyperlinks in the SOP should make it clear to the reader exactly where the other detail may be found. For example, if representativeness and target populations are fully explained in the protocol narrative or in the chapter on sampling design in the plan, then the representativeness section of the QA/QC SOP should clearly "point to" the section where the subject is fully covered. Details related to individual sites and individual measurements or parameters ("characteristics" in STORET terminology) should usually be fully explained in the representativeness section of the QA/QC SOP.

To obtain data comparability, it is OK and even desirable to use well established QA/QC procedures of another federal agency (USGS, NOAA, EPA, NAWQA, or EMAP) or a state agency. In the QA/QC SOP, list the source-agency, measurement quality objectives, and SOP details for sensitivity/detection limits, precision, systematic error/bias and blank control (the latter for chemical labs only). The SOP source-agency (EPA, EMAP, USGS, etc.) may change their SOPs as time goes along, and we need to have solid documentation of the methods we used at the start. For subsequent method changes, see section below entitled "Archive Cumulative Bias from Method Changes in the Data Analysis SOP".

In some cases, the same QC measurement quality objective (MQO) can be given for several parameters in a suite of vital signs included in one protocol. For example, if a network decided to use EPA marine EMAP QC SOPs to obtain maximum data comparability with EPA and state marine EMAP data, they could specify a precision repeatability MQO of 10% for several parameters to be measured in the field, including pH, temperature, DO, specific conductance, salinity depth, light transmittance (PAR), turbidity, and Secchi depth. The EMAP methods are a good source of MQOs for precision, bias, for field probe measures (EPA. 2001. Environmental Monitoring and Assessment Program (EMAP): National Coastal Assessment Quality Assurance Project Plan 2001-2004. United States Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Gulf

Ecology Division, Gulf Breeze, FL. EPA/620/R-01/002, http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf).

However, in many other cases MQOs will be different for different parameters and can simply be listed in a QC SOP table. A protocol for water column parameters measured in the field would typically have different measurement quality objectives than a protocol for nutrient parameters measured in the lab. However, in both cases, a table in a separate QC SOP in each protocol could list the MQOs for each applicable parameter.

Be careful with QA/QC terminology. Words and phrases such as representativeness, sensitivity, detection limits, accuracy, precision, repeatability, reproducibility, error, systematic error/bias, and uncertainty, have been used for different concepts by different groups. The confusion in water quality and contaminants QA/QC terminology has been so widespread that it brings to mind a "Tower of Babel" (everyone speaking different languages, no one understanding each other) scale of confusion. Some care has been taken herein (and in more detail in Part B) to explain the right terminology and to standardize on National Institute of Standards and Technology (NIST) and International Organization for Standardization (ISO) terminology wherever possible.

If the QA/QC of another agency is not adopted, or if the QC details come from multiple sources or are brand new, before completing the QA/QC SOP, a final check should be made to make sure that the measurement process will be controlled in some documented and defendable manner. The networks need to document what will be done for each of the issues listed below (doing nothing is not an option), but the networks need not "go overboard." However, at minimum, reviewers will be looking for common-sense documentation related to each of the following QA/QC basics:

Monitoring Design, Representativeness, and Target Populations

These QA topics are discussed together, because the way one assures representativeness is to name the target population and then design the monitoring to sample the target population in such a way that the samples obtained: 1) are representative of the target population, and 2) help answer previously identified questions (see discussion above). By the time networks are working on protocols and SOPs, many basics relevant to these topics should have already been summarized in chapter 4 of the central monitoring plan. The new task is to put additional detail about target populations and how representativeness will be assured in each protocol narrative.

Target and/or sampled populations can be summarized in a table in the protocol narrative. Additional detail on exactly what will be done to ensure representativeness should be place in the representativeness section of the QA/QC SOP. The planning process is typically iterative with continual refinement, so when changes are made in the protocol narrative and SOPs, go back and make sure all of the related discussions in Chapter 4 of the central monitoring plan are still correct and consistent with the additional text.

For example, suppose a network initially named a target population as "flowing waters of the network." Suppose further that sampling was in fact only going to be done in the daylight in low flow conditions during a July and August index period, and only riffles were going to be sampled. The sampled population (and sphere of inference and

conclusions) then includes only those potential values that could be measured during those very specific conditions.

Representativeness

Representativeness is a quality assurance basic that needs to be discussed in a complete manner in every QA/QC SOP. It also needs to be discussed in less detail in the protocol narrative. Together the protocol narrative discussions and SOPs need to answer the following questions: "Representative of what?", and "How will it be assured that the samples taken are representative of the target or sampled population?" These questions need to be answered even if USGS or other widely used protocols are utilized.

DOI information quality guidelines as well as more generic QA/QC guidance documents encourage "a high degree of transparency." One reason that defined target populations or sampled populations are compared to sampling designs and questions to be answered is to help insure transparency. In other words, don't hint that your conclusions apply to all waters of the park when they really are only applicable to daytime only, late summer only, low-flow conditions only, riffles only, one stream only, or near one specific bridge only.

There is growing recognition that unless care is taken to ensure representativeness, data can be of little value, no matter how good the measurement performance is for precision, bias, detection limits, etc. In other words, ensuring data quality means not only insuring analytical quality but also sample representativeness of the target population given the questions to be answered (see Part B).

One typically ensures representativeness statistically by having defendable monitoring designs, typically incorporating at least some randomness (as is suggested in both generic Vital Signs monitoring guidance documents and in Part B, see the later for more detail).

Given what is known about variability in time and space, how will the sampling scheme insure that the values obtained will be representative of the target population being monitored (generic VS checklist, op. cit.)? If the answer is not in the protocol narrative, a statement should be made in the narrative as to where to find the answer. As one hypothetical example, the protocol narrative might state:

> "Twenty five to fifty stratified random samples (or GRTS-selected samples) will be collected from all flowing waters in the park. All flowing waters of the park, at all times of day and times of year and all locations, will have a chance to be selected, assuring representativeness to all flowing waters of the park. The target population and the sampled population are both "all flowing waters of the park" (see more detailed discussion in the representativeness section of the QA/QC SOP attached to this protocol)."

Although this would be a good example, most monitoring networks would probably reject something this broad due to cost and other practical considerations. A network might start with a very broad definition of the target population when first writing chapter 4 of the central monitoring plan. However, later when they start further

developing protocols, SOPs, and monitoring plan optimization steps (see Part B), they would probably then whittle it down to something more realistic (see more realistic examples for copper and arsenic below).

Target Populations versus Sampled Populations

The target population is simply the larger universe of all possible values (bounded in time and space) that one is sampling from and wishes to make statistical inferences (conclusions) about. This definition assumes the ideal situation where the target population and the population actually sampled are the same. Note to biologists: the "target population" usually does not necessarily mean a biological population in the sense that biologists often use the phrase, as a specific level of organization (contrasted to the higher "community" or lower "individual" scales of organization).

Many monitoring and statistical guidance documents state that a target population and a sampled should ideally coincide (see Part B for more details). For example, an OMB committee came to this conclusion and also stated that if there is a large set of units in the target population that has no chance of selection, the design is not a probability survey (Federal Committee on Statistical Methodology, OMB. 1988. Statistical Policy Working Paper 15 at http://www.fcsm.gov/working-papers/wp15.html).

Most monitoring networks cannot afford to randomly sample all habitats at all times and in all places. Therefore, it is often useful to initially define target populations in very general terms (say all waters or all flowing waters or the park or network) and then later specify a more restricted "sampled population," with inference only extending to the sampled population.

If a monitoring networks decide to make a distinction between target populations and sampled populations, (as recommended by EMAP at http://www.epa.gov/nheerl/arm/surdesignfaqs.htm#whatpopulation), then it is important to define both in as much time and space detail as possible. If no sampling is to be done in the winter or at night, the sampled population and sphere of inference or conclusions should not include night time or winter conditions.

In final reviews about whether or not what is proposed makes sense, networks need to compare the sampled and target population to the questions to be answered and extent of inference.

Copper Example:

For example, suppose the question is "does the concentration of copper in the water column in all flowing waters of the park ever exceed state chronic water quality standards for aquatic life?" In this case, the network would typically need to consider when worst case conditions typically occur. If the network really wants to determine if there were ever any exceedances of the standard in any flowing waters anywhere in the park, then one would not want to restrict sampling to riffles only. Furthermore, if one wanted to know if copper standards were ever exceeded, then one would sample at night, the time where water column samples of copper tend to be the highest, not during the day. Fish and other aquatic life do not just live in the water during the day.

To answer this question, one must consider worst-case conditions. One would restrict sampling to short summer index period (for example, July and August low flow periods), if one already had reason to believe that is when copper concentrations would be the highest.

After considering the above example for copper, the network might realize that it was unrealistic considering budgets available and decide to make some new monitoring plan optimization changes in 1) target or sampled population, 2) extent of inference, and 3) the basic monitoring design. If the network really wanted to answer the question of copper standards exceedance, the discussion in the protocol narrative might include the flowing example.

1. The target population might still be "waters of the park," but only if the network also defined a "sampled population" and only if the sphere of inference and conclusions are not to extend beyond that sampled population.
2. In the study design part of the protocol, the network might clarify that were stratifying by time of year, by flow conditions, and by night-only times for sampling. If the network or their advisors prefer not to call these restrictions strata, they can simply call them "response design details."
3. Again, for emphasis, no matter what terminology is used, the key factor to be stated in the protocol narrative discussion of target populations, representativeness and monitoring design, is that sampling will, in fact, be done only during those restricted spheres of time and space, and that the extent of statistical inference (and the sampled population) will not extend beyond those restrictions.
4. The type of probabilistic sampling design should be listed in the protocol narrative (random, stratified random, or GRTS). The type of design could be lined up with questions and basic approaches in a table in the protocol narrative. Some sort of probabilistic design would be needed if inference were to be made broadly beyond the areas and times immediately sampled.

Budgets often restrict the sampled habitats, sampled locations, and/or sampling times. The sampled population would seldom include all waters of the park (big rivers, small wadeable streams, lakes, wetlands, and ponds). Indeed, different protocols & SOPs would typically be needed for each of these radically different types of habitats.

Refine Target Populations, Questions to Be Answered, Sampling Design and Representativeness in an Iterative Manner

As an example of how monitoring planning often proceeds in an iterative manner in the gradual fine tuning of monitoring details, let's consider what the next step might be. The network might reconsider some of the details listed above. Would the network really be able to do the night time sampling on a regular basis? Is that really the most important question to answer? The network might decide that it was just not realistic to sample at night. They might consider answering a different question instead, considering the limited monitoring funds available and other real-world practicalities.

Such changes are not unusual when monitoring networks are optimizing monitoring designs. Often monitoring networks have more vital signs, measures, and questions than they can answer with the budget at hand.

Again, optimization steps usually include throwing out vital signs, measures, or questions. Some may be thrown out because they require night time sampling or other aspects considered impractical or dangerous. Others might be thrown out because some other agency is already covering the monitoring. Still others might be thrown out due to excessive variability (even at pristine sites) and a resultant inability to fine trends against the background of the high variability.

As an instructive hypothetical example, let's suppose that not only is copper a potential issue at park, but so is arsenic. Suppose further that another equally high priority question was "Do water column concentrations of arsenic flowing into our linear (riverine habitat) park ever exceed water column samples for arsenic?" That would be an easier question to answer, because water column values of arsenic tend to be highest in the afternoon rather than in the middle of the night. Also, only one site would need to be monitored. The protocol narrative discussion might then be changed to reflect the following:

1. The protocol narrative would state that a "targeted" ("judgmental") sampling design is appropriate to answer the question rather than probabilistic or random design. The question does not relate to all waters of the park and to answer it we only need to sample where the river flows into park jurisdiction
2. Likewise, the target population is no longer "flowing waters of the park." Now it relates to only one location. The sampled population might be "water flowing into the park, where the river crosses into the park boundary (or close)."
3. In this hypothetical example, let's assume that the network has no knowledge of one season being worse than another. The protocol might then state that 30 sampling dates (during the course of a year) will be picked randomly. If arsenic is worst case during a narrow seasonal index window of time, the protocol narrative should state that and specify monitoring will be done within that window of time.
4. To capture worst-case conditions for arsenic, sampling of the water column would be done in the afternoon only, when arsenic was likely to be highest in the water column.
5. The sampled population and extent of inference would therefore not extend beyond afternoon conditions. Also, since sampling will only be done in one location (where the river comes into the park), the extent of geographical statistical inference will not extend beyond that one location
6. In the study design part of the protocol narrative, the network might clarify that were stratifying by time of day to try to take out most of the diel variability (most water column parameters show diel variability, especially pH, oxygen, temperature, chlorophyll, metals, and nitrates, see Part B for details). Or they could simply clarify that sampling will be done in the afternoon only, as part of response design SOP details.

If the network in our hypothetical example decided that they want to keep an eye on copper trends, even though they can't practically sample at night, they might decide to

look at trends rather than water column exceedances. They might further decide to sample copper in sediments rather than the water column. If they understood local variability enough, response design details might call for other sampling restrictions. For example, in one small stream in Yellowstone, it was discovered that variability could be reduced by sampling metals in sediments only in low flow late summer conditions and only in the sediments of low gradient riffles, where variability is lower than in the water column or in other sediment microhabitats (such as backwater pools).

Trying to get the variability down is done so that the monitoring network can have a reasonable chance to detect a change of concern (say a 30% change in means over 10 years) without collecting hundreds of samples every time they went out. See Yellowstone example in Part B.

Some networks (and many states) use hybrid sampling plans that include both 1) targeted sites (such control sites or historically sampled bridge sites) to answer site-specific or other limited inference questions and 2) probability-selected sites that allow for broader inferences to larger areas of the park or waterbody. Such hybrid designs are often good compromises but sometimes tend to stretch funding even further and make getting 25-50 samples per year in each park more difficult.

At least two networks (SECN and PACN) haven tentatively proposed to solve the problem by getting 30-50 probabilistic samples per year at one park only, and rotating to other parks in future years, while using continuous monitoring at targeted/judgmental sites to cover both impacted and pristine fixed sites at each park each year to better understand temporal variability.

If All Sites Were Selected With a Judgmental Approach

If absolutely no randomness is be involved in picking sites to sample, the rationale should be justified on a network-specific basis. If no probabilistic methods are to be used, why not? With what logic would a targeted design assure representativeness and why do the pros of the targeted design outweigh the cons, given the money available, or the need to continue historical trend data?

If a network chooses to make all sites judgmental or targeted sites, with absolutely no randomness at all, they still need to address representativeness and target population. In the absence of convincing evidence to the contrary, the target and sampled populations, as well as the extent of inference, will all be limited to those sites sampled only.

Even if the stream near a bridge is selected in for long term monitoring, there are things that can be done to approve the quality and usefulness of the data. These might include:

1. Sampling far enough way from bridges to minimize bridge-effects (deicer salts, dust, vehicle pollutants, trash, changed hydrology, etc.) to help with representativeness with this general area of the stream and not just the (often unusual) conditions right at the bridge.
2. Once the network has picked the area to sample (say upstream of a bridge, possibly in riffles only), they can still pick exactly where to sample in the riffle randomly, using simple stop watch field randomization (see Part B).\

Alternatively a network may decide to use guidelines such as those used by the USGS National Water-Quality Assessment Program (NAWQA) sediment quality collection protocol to maximize data comparability with NAWQA. That protocol (http://ca.water.usgs.gov/pnsp/pest.rep/bs-t.html) specifies collecting sediment samples in low flow periods only (to reduce seasonal and flow driven variability), and they specify compositing samples from different microhabitats (within and among different zones) to get an average for a reach and to reduce variability driven by habitat type (and to make the samples more representative of a larger area). Other parts of USGS and EPA have also used similar types of guidelines for obtaining representative samples. For the water column such guidelines often involve compositing cross sections, determining if centroid samples are representative, and the like.

Using such guidelines is fine and often has the advantage of helping achieve data comparability with other agencies with other large regional data sets. Again, one still has to address the question: "representative of what?" Whatever understanding one has of terms like strata, the target population, the sampled population, and zone of statistical inference, all such phrases should be clearly defined and be transparent to readers.

If the agency believes that sampling a cross section of surface water will be representative of some areas upstream that have not been sampled, how far upstream, and based on what data? At minimum, comparisons should be done before starting monitoring to check all such assumptions, and these should be repeated occasionally over the years. Otherwise, such beliefs will be based strictly on speculation rather than on any actual data. It would take many such comparisons (more than most networks can afford) to convince most statisticians and survey design experts. Thus the most common practical alternative is simply to state that inference will not extend beyond the particular site location (and/or times) that had a chance to be monitored.

Causation

Documenting causation (not a requirement in VS monitoring) is difficult to prove without active manipulation. Inside labs, only one variable is typically changed at a time, and the rest kept constant. Outside in the environment, countless variables (temperature, sun energy, wind, etc.) are changing all the time, often in unknown ways. So to get at potential causation, one usually uses multiple lines of evidence approaches, such as those explained in EPA's 2000 stressor identification guidance (http://www.epa.gov/ost/biocriteria/stressors/stressorid.pdf). That document includes plain language summaries of how to do strength of evidence analyses using multiple lines of evidence.

None of the above prevents monitoring networks from thoughtfully placing sites in such a manner that hints relative to causation can be obtained, but remember that stressors tend to change over time and this is long term monitoring.

Stratification

If a monitoring network has decided to stratify, they should avoid using strata where variability characteristics (in time and space) are not well understood or are likely

to change appreciably during the monitoring period. Keep in mind that this is long term monitoring, so eventual change is more likely than for short term projects. It is often safer to stratify by factors that change less frequently or dramatically (often geological or physical factors), or to handle timing and detailed space issues in the response design rather than in a more general monitoring design in chapter 4 of the central monitoring plan.

Typical patterns of variability and typical patterns of response to various single stressors say nothing about other patterns of variability and other responses not considered when trying to determine "typicalness." In other words, it is often harder to group sites into homogeneous groups, especially for multiple stressors, than one first thinks (http://oregonstate.edu/instruct/st571/urquhart/represent/sld028.htm).

Often monitoring planners need to think through collection details carefully in order to get variability down to magnitudes that would allow detecting a reasonably small change without hundreds of samples, For streams, if variability characteristics are understood well enough, one can stratify by habitat types (such as low gradient riffles only, or runs only, or snags only). Monitoring groups are sometimes able to document that the variability reducing aspects of stratification outweigh the disadvantages. The stratum description might be qualified to take into account the changing environment of streams, climate change, etc. One can also specify index collection time periods in either stratification decisions or in response design decisions.

A useful document on the typical need for stratified random sampling of outdoor environments (classified as non-experimental studies of uncontrolled events) was provided by Schwarz 1998 (Chapter 3 in Statistical Methods for Adaptive Management Studies.

Again, if a monitoring network chooses to handle such details under the response design documentation rather than calling it stratification or handing them in stratification steps (as part of the spatial monitoring design), one can put the needed details in individual protocol narratives and SOPs.

GRTS and Similar Approaches

To obtain the logistical, safety, spatial balance, and other benefits of unequal weighting, many networks have chosen to use generalized random-tessellation stratified (GRTS) design rather than more simplified versions of stratified random sampling. Both GRTS and stratified random sampling involve probabilistic strategies. No matter how such issues are decided, all decisions should follow a careful and documented (in the protocol narratives) thought process,

GRTS allows inference to a broader target population and broader are is the generalized random-tessellation stratified (GRTS). GRTS can be used with our without stratifications and various monitoring networks have decided to use GRTS.

Another solution and variant on spatially balanced probability design is the Systematic, Unequal Probability Sampling Design planned by the Northern Colorado Plateau Network (NCPN).

A combination of unequal weighting of probability of selection (plus, in some cases, stratification to reduce variability) can result in more intensively sampling certain targeted sub-regions and help negate the likelihood of getting sites where personal safety

or access are a big problem. There are also ways to accommodate existing monitoring sites (with historical records) within a GRTS design. The issue is one of assigning an appropriate weight to the non-randomly selected points, i.e., figuring out what fraction of the population is represented by a non-random point. Spatially-balanced samples are random samples. They just don't happen to be simple random samples or systematic grid samples. For more information on unequal weighting, spatially balanced designs, GRTS, etc. see http://www.epa.gov/nheerl/arm/designpages/design_tech_info.htm.

GRTS and other probabilistic designs in general are often good and often even superior to other schemes, when done right, when sample sizes are high enough, and when all aspects are logically defendable. Again, the big draw to such designs is the ability to infer to larger areas and not just to those being sampled. Many such designs are designed to produce proportions (% of stream miles impaired, % of flowing water achieving water quality standards, % of flowing waters where an index results in a rating of excellent, etc.).

For those with quantitative backgrounds, more information on GRTS as applied to Vital Signs monitoring can be found in the related presentations at the San Diego National Vital Signs Meeting in 2006 (See presentations by Schweiger and Urquhart at http://science.nature.nps.gov/im/monitor/meetings/SanDiego_06/SanDiego.cfm).

GPRA and Proportions

Proportions are potentially useful for GPRA and other management and reporting goals. One caution: Beware of (or at least look closer when encountering) small sample sizes when estimating proportions. Keep in mind that sample sizes should be 30-50 to estimate a proportion well and that any proportion estimated for sample sizes below 25 is a big red-flag (see EMAP "Why a sample size of 50" detailed explanation at http://www.epa.gov/nheerl/arm/surdesignfaqs.htm#manysamples). See additional discussion in the proportion section of the sample size calculators (below, under general heading of completeness).

A typical problem for NPS VS networks is that they often cannot afford (the optimally defendable for a proportion estimation) 50 aquatic samples per year in a GRTS design unless they use other generic VS funding to supplement water quality funding. Like many states, many networks do not want to put all their funding into a GRTS design but instead they often favor a hybrid design. Networks often desire to monitor at least a few targeted sites for long term continuity or to answer site-specific or resource-specific questions. Most networks also want to measure more than one aquatic variable and/or different variables at different types of sites.

Does It Still Make Sense?

A potential complication encountered by other networks using GRTS or other probabilistic designs, has been that lumping values from different years together to eventually get a big enough sample size may not always be logically defendable. On one hand, lumping five years of data might help cover a fuller range of conditions better than single years.

On the other hand, very small sample sizes (always problematic or at least worrisome in statistics) can be a fatal flaw, especially when combined with inattention to timing and spatial issues. If one only takes 30 samples from a very large area (such as a whole park or whole network) over one five year period, could one stand up in court (or even in front of a superintendent) and say with a straight face that the full range of conditions had likely been captured with our 30 samples? Although the fact that samples had been sampled randomly is perhaps even more important than sample size, having a reasonably large sample is also crucially important.

In the above example of 30 samples, the resulting proportion based on five years of lumped data may not be fully or optimally representative of the target population one was trying to protect. What would the target population or sampled population be? Whatever we do should pass last minute reality-checks of logic, defendability, explain-ability (keep it simple is optimal there), and common sense.

In the highly variable universe of water quality, it may be hard to logically defend the notion that five annual samples, of sample size six each (from a very large area) is in fact "one sample" and the right (or optimal) sample to estimate a proportion or average. Due to changed conditions, it might be easier to defend that there are in fact, five valid samples (not one). In worst case scenarios, GRTS plus excessively small sample sizes and inattention to timing and important variability-reducing response design details (such as sampling low gradient riffles only or snags only, or only in short "index" time periods) may produce data that is so variable and so anecdotal (small sample sizes) that it may not useful for many (if any) purposes. This is the very reason why so much water quality data collected in the past has not been useful for management purposes. Another complication with compositing is that it complicates power analyses. This is not a fatal flaw by itself, but must be dealt with in defendable ways (see additional discussions below).

An important reason past data has too often not been useful relates to trend detectability. Five year averages estimated from extremely variable data might make it difficult to detect even big changes from one five year period to the next, or to detect longer term trends. Again, our monitoring design should produce data that is useful for resource management decisions and useful to answer stated questions.

Timeliness is another issue to consider. Resource managers may not consider conclusions that they get only once every five years to be timely enough. Superintendents have sometimes wanted to detect a change of less than 50% over one year. For certain rare or important biological resources, superintendents have sometimes not wanted to lose 50% without knowing it after one year, let alone after five years. This may be even more worrisome if the estimate is questionable due to small sample sizes.

Reporting data and QA/QC summaries (but not conclusions on trends or water quality exceedances) annually is necessary and helps. However, if after 5 years, meaningful summary statistics (means, medians, proportions, water quality standard exceedances, % meeting acceptable condition index scores, etc.) cannot be calculated, that would typically be a problem. Likewise, if after 10-20 years, if even true and substantial trends could not be detected because of study design flaws (often including inadequate sample sizes), resource managers and other data users will probably not be well served. Species which are legally protected or even locally rare would be difficult to manage based on conclusions once every 5 years. Again, they might disappear between conclusions. In such special cases, there may be missed opportunities for management,

and resource managers may have very little warning about declining populations. With shorter intervals of monitoring, and credible sample sizes, there is a better probability of detecting trends or bad conditions in time to develop and implement management strategies to avert losses.

The SWAN Phase III draft addresses some related "common sense" tests in rotating panels in a bit more technical terms: "An important consideration when choosing a revisit design is its ability to retain a representative sample across time. A sample that is initially representative may lose this quality if there are changes or shifts in population numbers or other attributes during later time periods that are no longer captured by the original sampled units. These shifts across time could be induced by natural changes (e.g., habitat succession), anthropogenic actions, or a combination of both. If large shifts are not expected to occur or if the membership design is spatially balanced enough to adequately capture any shifts, the best revisit design to detect trend is to repeatedly sample the same plots across time, all else being equal. However, repeated visits to the same units could potentially have a negative impact on the response, such as trampling in vegetation monitoring plots, which would introduce bias" (http://www.nature.nps.gov/im/units/swan/Libraries/MonitorPlan/MP_Phase_III/1_SWAN_2005_MonitorPlan/BennettA_2005_SWAN_MonitorPlan_051215.pdf).

Will the Sampling Design Produce Information Useful to Management?

A key question is what kinds of data would be of most interest to management? Would a superintendent really be more interested in results from a sampled-population stretched over five years, or would that superintendent be more interested in how things conditions are in wet years vs. really dry years? Wet years might result in different results (different means, different variances, etc.) than dry years. To protect the resource, superintendents may need to manage the resource differently in wet vs. dry years. Therefore, the superintendent may be more interested in defendable estimates of what happens in those two conditions than in what happens over 5 years with many different conditions.

Another related issue relates to internal data comparability. Are the data sets from five different years comparable enough to be combined into one sample? Over five years, biotechs and other personal (and equipment) are likely to change, resulting in measurement bias changes. So logically one might then have two different samples, samples where biotechs measured higher, another sample where biotechs measured lower. Variances may also change for similar reasons. Both variances and means would also be apt to change due to changes in condition and true variability in the population being sampled. These types of changes would be more apt to happen over a five year period than in one season, and would make it harder to defend that we logically have only sample and not more.

States that require relatively few samples (for example, 1 per month, or 4 per year for metals, both for two years) for the purpose of gauging compliance with water quality standards may be exceptions. In that case, there are enough samples for regulatory compliance simply because the state says that is enough. However, even then the

resource manager may recognize the need for more samples thoughtfully placed in space and time. For example, given that many metals vary diurnally, seasonally, and spatially, are 4 metal samples per year in a stream or reach logically enough samples to ensure scientific credibility (for example, to represent the full range of conditions in a representative way)? Usually not, and resource managers are usually interested in the true conditions and not just regulatory status. Only when the true condition is known can resource management be done in an optimal way.

Completeness, Sample Sizes, Statistics, and Detection Probabilities vs. Desired Conditions

Completeness is usually considered a QC topic, but to assure completeness, one must first consider sample sizes, overall monitoring or survey plan, detection probabilities, desired conditions and some other QA factors that are usually first mentioned in the central monitoring plan.

In aquatic Vital Signs Monitoring, Data completeness goals are typically given as percentages in tables in the QA/QC SOP or QAPP and are developed by first estimating required sample sizes. Although written for aquatic and water quality monitoring, a statistician who reviewed the following section reminded us that most of these steps are generic and would also apply to terrestrial monitoring.

Determining required sample sizes and attendant completeness goals should be done in a stepwise manner, considering the following in a more detailed and quantitative way than has been done in earlier planning phases:

1. Refine (provide more time and space detail) objectives and questions
2. Identify desired conditions qualitatively.
3. Identify resource-collapse or other thresholds (such as water quality standards or no-effect levels)
4. Identify existing conditions.
5. Develop safety margin between existing conditions and threshold magnitude.
6. Document variability in time and space.
7. Refine target population details.
8. How big of a difference or change do we need to be able to detect?
9. What initial statistics will be used?
10. Choose desired detection probability/statistical power (1-beta).
11. Choose statistical significance level (alpha).
12. Use simple calculators to make initial estimates of required sample sizes.
13. Throw out measures or strata where excess variability will prevent detecting a trend or a difference of concern within budget.
14. Optimize monitoring plan details for affordability and logic.
15. Draft initial sample sizes and optimized monitoring design.
16. Finalize sample sizes and design with an applied environmental statistician.
17. Estimate the % of samples that will fail (for example 10%).
18. Increase the planned sample sizes accordingly.
19. Put completeness goals in a table in the QA/QC SOP.

The first three above are typically covered in varying degrees of detail in the central monitoring plan. Some have also previously been introduced herein (above) in the section on Objectives and Questions. However, when developing the fine details of the monitoring design, sample sizes, and statistics, several of these inter-related issues should be reconsidered in a more thorough and quantitative way and documented in more detail in each protocol narrative and in relevant SOPs. The goal would be to make sure they all line up and make sense when considered together. Defining the first two steps in as much time and space detail as possible is helpful when moving to the more quantitative steps (3-19).

When faced with a 19 step process, why not just go to a professional statistician to start with (or maybe starting along about step 4)? Great idea, if the network can afford it. However, many of the steps are decisions to be made by the park or network, not the statistician, and would in fact be input to bring to the statistician. All of the steps before 16, except perhaps 9 and 12, should be done by NPS staff, often with the help of the network quantitative ecologists. Even if performed by a statistician, the statistician would need considerable input from the NPS in going through the steps. Furthermore, bringing Vital Sign network quantitative ecologists up to a certain minimum level of understanding is a good goal and one that would help prevent some past disconnects between the statistician's advice (often in Chapter 4 of the central monitoring plan and mistakes made later by networks in protocol and SOP development after the monitoring staff stopped talking to the statisticians.

Fully informed quantitative ecologists can help park management refine the steps above in an adaptive management way (see Statistical Methods for Adaptive Management Studies). For example, after step 15 above, it may become clear to all that initial decisions made for steps 1, 5, 6, 8, and 11 have to be adjusted for the design to make sense and be within budget.

Determining required sample sizes and data completeness goals admittedly takes a bit of effort. However, is especially important for long term monitoring and failing to do so has all too frequently resulted in aquatic monitoring that has produced data that has not been useful for management decisions. Too often raw data has never made the transition to useful information ("Ward, R.C., Loftis, J.C., and G.B. McBride. 1986. The "data rich but information poor" syndrome in water quality monitoring. Environmental Management 10:291-297).

A bit more detail on each of the outline steps is provided as follows:

1) Refine (Provide More Time and Space Detail) Objectives and Questions

Why revisit and refine questions? The monitoring design and statistics to be used are both driven by the questions to be answered, and it helps if the questions to be answered (and the identified target population being monitored) are as detailed in time and space specifications as possible (see earlier sections on Questions and Objectives and on Representativeness and Target Population). Again, it is very important that all of the concepts in the following outline line up and be reasonable when all are considered together.

If one calls them objectives rather than questions, the details of what, where, when, and (even) how big of a change can we detect; all still need to be detailed before

one can design monitoring in an optimal way. The number of objectives competes with the number of samples in a cost-limited study (Kurt Jenkins, USGS BRD helping NCCN with birds, Personal Communication, 2006).

2) Identify Desired Conditions Qualitatively

The central monitoring plan should document (to the degree possible) desired future conditions (DFCs, hereafter abbreviated to desired conditions or DCs). At the protocol development stage, additional detail on DCs should be placed in the protocol narratives. For a generic (not just water) Vital Signs monitoring discussion of DCs, see talks by Steve Fancy and Rob Bennetts at http://science.nature.nps.gov/im/monitor/meetings/SanDiego_06/SanDiego.cfm. Some of the key points therein and other related points are summarized briefly as follows:

In the new-style NPS General Management Plans, DCs are defined as "A **qualitative** description of the integrity and character for a set of resources and values that park management has committed to achieve and maintain"(NPS, 2005. appendices and glossary portions of the revised NPS Planners Sourcebook at http://classicinside.nps.gov/documents/4%20OCTOBER%20APPENDIXES.pdf, main document at http://inside.nps.gov/waso/custommenu.cfm?lv=2&prg=50&id=3317).

Additional insight may come from the new-style Resource Stewardship Plans (RSPs), Park Superintendents, and Park Resource Management Staff.

The NPS Management Policies notably emphasize both protection of park resources for the enjoyment of future generations and sustainability of both natural and financial resources (http://parkplanning.nps.gov/projectHome.cfm?projectId=13746).

When considering desired conditions, it is good to be "realistic" by recognizing that desired conditions, or even possible conditions, tend to be moving targets (and to vary within fairly broad ranges) due to continued changes in climate, invasive species, direct and human impacts, and countless other changes in stressors upon which resource managers often have limited control. Emerging concepts and complications include applied historical ecology, the fragmentary nature of history, the subjective and value laden aspects of desired condition goals, and pre-Columbian impacts of man. Other issues include assumption, difficult-to-quantify confounding factors, no-modern-analog issues, and non-equilibrium paradigms. It is often fruitless to choose a single fixed point goal and better to use ranges (Swetham, T.W., C.D. Allen, and J.L. Betancourt, 1999. Applied historical ecology: using the past to manage for the future. Ecological Applications 9(4):1189-1206, http://www.fort.usgs.gov/products/Publications/258/258.pdf).

It is also appropriate to think through where we are now and how big of a change from the current condition would it take to move beyond a negligible, minor, or moderate impact as defined in NEPA terminology (NPS 2003 Interim Technical Guidance on Impairment of Natural Resources (http://www2.nrintra.nps.gov/ard/docs/nrimpairment.pdf). Therein, parks decide whether or not a **biological** impact is negligible according to the following definition):

> **Negligible Biological Impacts:** Impacts occur, but are so minute that they have no observable effects on plants and animals and the ecosystems supporting them. The severity is "Trivial effects on individual organisms or areas of habitat." The

duration is “Short-term to long-term effects.” The timing is: “Outside of critical timing windows of key resources or ecosystems.”

Parks decide how big water quality impacts are according to the following criteria):

Negligible Water Quality Impacts are described as “Impacts are effects that are not detectable, well below water quality standards, and within historical baseline water quality conditions.” The impairment guidance (op. cit.) also describes other levels of water quality impacts:

Minor: Impacts are effects that are detectable but well within or below water quality standards and within historical baseline water quality conditions.

Moderate: For most waters, impacts are effects that are detectable, within or below water quality standards, but historical baseline water quality conditions are being altered on a short-term basis. However, in outstanding natural resource waters (ONRWs), this threshold may approach the requirements for statutory impairment.

Major: For most waters, impacts are effects that are detectable and significantly and persistently alter historical baseline water quality conditions. Water quality standards are locally approached, equaled, or slightly and singularly exceeded on a short-term and temporary basis. However, in ONRWs this threshold would probably constitute statutory impairment.

In highly pristine resource parks or parks with highly valued, rare, or endangered resources will DCs be defined as negligible impacts according to the above? How will qualitative goals be translated into quantitative goals?

If the water is already much cleaner than default state water quality standards, will stronger quantitative water quality goals or anti-degradation standards be used?

At the other end of the spectrum, if one is considering a historical park in a highly urbanized, industrialized, or farmed area where there is no reasonable expectation of ever achieving negligible impacts, the area may be in an alternate steady state. In this scenario, the DC may simply be avoiding additional significant impacts, or perhaps meeting urban habitat state water quality standards.

Initial desired condition (DC) goals need not be permanent. Gradually refining DC goals in a classic DOI-style adaptive management cycle might follow a pattern such as this (expanded slightly but based on suggestions by Steve Fancy, February 06 at hhttp://science.nature.nps.gov/im/monitor/meetings/SanDiego_06/San%20Diego%20DFC%20Breakout.ppt#262,6,Slide 6):

1. Park staff qualitatively describe the DC in the new General Management Plans,
2. Park staff compare current conditions to DCs. Decisions should be documented in the new cycle of developing Resource Stewardship Plans. At this stage, parks are

perhaps still qualitative to some degree but are at least beginning to develop or think about quantitative performance measures,

3. Park staff members develop and implement management strategies to achieve desired conditions. 4) Park staff (at times possibly supplemented by Vital Signs planning and monitoring information already completed) finishes developing quantitative performance measures for monitoring. At this stage it would be optimal to be as quantitative as possible and define minimum detectable differences of the monitoring design,
4. Monitoring is done to detect trends in resource conditions and evaluate management effectiveness (see BC website at http://www.for.gov.bc.ca/hfd/pubs/docs/lmh/lmh42.pdf Or http://science.nature.nps.gov/im/monitor/docs/BC_LMH42.pdf),
5. Park management uses the results from monitoring to take adaptive management actions to help achieve desired conditions,
6. Park staff returns to (1) above, and refine DCs if appropriate, and
7. Park and monitoring staff keeps repeating the cycle and refining each step as appropriate as lessons are learned.

3) Identify Resource-Collapse and Other Thresholds of Concern

Park resource managers would typically desire to detect a change smaller than one that would cause a resource collapse, to provide a safety margin.

The monitoring protocol narratives should identify a resource collapse threshold for each vital sign or measure, if one is known, but often resource collapse thresholds are not known. However, one advantage of long term monitoring is that our understanding of resource dynamics and thresholds and threshold models can be refined in an adaptive management fashion as more data is collected. There are many examples of this in fishery literature (for example see 2004 NOAA Surfclam and Ocean Quahog Quota Specifications at http://www.nero.noaa.gov/nero/regs/frdoc/2005specs_10-27-04a-web.pdf). A resource collapse in one location sometimes identifies the threshold and the threshold can then be used to protect against collapses in other similar habitats in the region.

When resource collapse thresholds are not known, one often uses water quality standards that already have a safety margin built-in.

Sometimes there are none available and one has to use other quantitative comparison benchmarks (see Part B for where to find them). As on example, the EPA Costal EMAP program has suggested additional regional coastal threshold criteria for many indicators (EPA et al., 2005, National Coastal Condition Report II, EPA 620/R03/002, http://www.epa.gov/owow/oceans/nccr/2005/Chapt1_Intro.pdf**).**

Various water, sediment, tissue, and soil benchmarks are also often used. Risk assessment-derived benchmarks, especially No Observable Adverse Effect Levels, No Effect Concentrations (NOAELs, NOECs), and Low Effect Levels, can be used as one starting point, as long as it is realized that they are often less than optimal since they are seldom based on considerable local or regional work or even sound statistics. However, in some cases one has to use what is available until something better and more defendable is available (adaptive management approach).

Summaries on data comparison benchmarks for metals and industrial organics and petroleum hydrocarbons in water, sediment, soil, and tissues, updated through 1997-1998, are summarized in the NPS contaminants Encyclopedia (http://www.nature.nps.gov/hazardssafety/toxic). The NPS encyclopedia contains general ecological toxicity profile information on 118 contaminants. For a listing of all 118 topics, see http://www.nature.nps.gov/hazardssafety/toxic/list.html and for references for all 118 topics see http://www.nature.nps.gov/hazardssafety/toxic/referenc.pdf.

Thresholds, including water quality standards, should be thought of in the context of "if-then" management decision rules discussed in section IV-C of Part B. If damage or toxic concentrations exceed such and such and magnitude, then we will……reduce visitation, begin remediation, reduce fishing pressure (or whatever is appropriate, see Part B).

4) Identify Existing Conditions

This is done by summarizing past data (see previous discussion of "Summary of Information from Past Data", above) as completely as possible. In some cases, pilot-scale studies will have to be done when little or no information is available.

5) Develop Safety Margin between Existing Conditions and Threshold Magnitude

How much does the ambient condition have to change to get to a threshold of concern or a specified desired future condition? If a resource collapse threshold is known, how much does the ambient condition have to change to get the threshold (the value below which the resource will not recover, or will recover at an unacceptably slow rate, also called a breakpoint by some)?

Again, if one is only 10% of the mean away from disaster, then obviously being able to detect an effect-size change of a magnitude of 20% of the mean is not good enough. Water quality standards usually have a safety margin built in, but many other comparison benchmarks do not.

Too often, monitoring projects fail to detect true anthropogenic effects (type II error, false negatives, the conservationist's risk) because of inadequate survey design. In studies that measure change, there must be a large enough sample size to detect the minimum effect, or smallest difference or change that will cause management action. The smallest change is usually defined in terms of an "effect size" or minimum detectable difference (MDD, see definitions in the text further below).

How is the effect size expressed? Is it in original units as a difference between a mean and a water quality standard or the difference in two means? This is a common definition and one used by EPA (http://www.epa.gov/owow/monitoring/calm/calm_appc.pdf).

Or is "effect size" the totally different concept used in behavioral sciences and recent water quality work, a percent of the standard deviation (difference divided by the standard deviation all times 100 to get a percentage), a percentage change in a proportion, or a model output function?

In the case of fisheries, thresholds have usually been defined following collapses, trial and error, and (sometimes) long recoveries. One reason for being precautionary is

that some collapses can be permanent, with recovery never occurring. If a species is on the edge of its range, or just making it for whatever reason, even a minor change, such as a climate change or new competing species, might prevent recovery. Or, as is the case for some endangered species, populations can sometimes simply become so small that they don't survive.

In one example that considered both collapse threshold (magnitude) and smaller (safety margin) effect sizes to be detected, a slow growing macro algae, *Hormosira banksii,* was found to readily recover from depletion down to 30% cover. Pilot studies indicated an average cover of 75-85% cover. To give some margin of safety, the critical effect size goal was determined; monitoring needed to be able to detect a 30% or greater reduction in cover. This would allow detection of a reduction of cover from 75 to 45% and would also provide an effect size a safety margin of at least 1.5 (45/30) times compared to the collapse threshold of 30% cover. The idea was to give management time to institute protective strategies well before the threshold of 30% total cover might be reached (Mapstone, B.D. 1995. Scalable decision rules for environmental impact studies: effect size, Type I and Type II errors. Ecological Applications 5: 401–410).

6) Document Variability in Time and Space

Whenever practicable, target population definitions and sample sizes estimates should not be finalized without estimates of variability in time and space. Calculating the initial probability of detecting an effect size in standard deviation units does not require input of a variance or nonparametric analogs. However, as soon as such estimates are available, they should be documented so that they can be used as input variables in estimating a minimum detectable difference (MDD, see step 8, just below) in original units of measure.

It is best to also have MDDs in original units, since they are more intuitive and understandable, and since standard deviations (SDs) can vary according the magnitude of signals and means. In other words, variability (as expressed by a SD) is often not uniform through the full range of different signal magnitudes. Exclude variability estimates if they are mostly based on signals not more than twice the MDL detection limits, since these would typically be much higher than for normal measurements (see part b discussion of low-level detection limits for more detail).

It is best if the initial variability magnitude estimates come from the areas to be sampled. If no such information is available, past data from similar habitats in the region may often be OK for initial estimates of variability. If there are no nearby data from similar habitats from past monitoring, pilot-scale monitoring may be needed.

Most of the simple and even more complex (simulation) sample size calculations depend on a good estimate of the SD, so be wary if the SD estimate is based on sample sizes under 30-50 unless variation is known (for sure) to be VERY small. Also be wary if the values used to estimate a SD do not cover the full range of time and space conditions of the identified target population. If the variability is very low and the full range of conditions has been covered, a few samples will be enough, but a few samples may be virtually worthless in the presence of substantial variation (the common case for most water column measures).

Stratified random samples, sometimes accompanied by narrow temporal collection index-period windows, are often used to reduce variability and make it easier to detect trends between years or over a period of several years. In some scenarios, GRTS can also be used to bring down variance. Rather complex ways to figure multi-year/multi-site variance can used for rotating panel designs. For those with strong quantitative backgrounds, more information on the complex ways to estimate variance can be found: 1) In additional references in Part B, 2) in the presentation by Urquhart at the San Diego National Vital Signs Meeting in 2006, which discussed not only variance aspects but gives examples of statistical power for various rotating panel revisit options (http://science.nature.nps.gov/im/monitor/meetings/SanDiego_06/SanDiego.cfm, and 3) in Stevens Jr., D. L., and A. R. Olsen. 2003. Variance estimation for spatially balanced samples of environmental resources. Environmetrics **14**:593-610. However, these discussions are more complex, than most of the related discussions herein, so talk to your applied statistician before finalizing an approach.

For contrast, simple random sampling or sampling at various times of year (or even various times of day) can 1) increase variability greatly for many water column or sediment quality parameters, and/or 2) greatly increase the number of samples needed, and/or 3) decrease our ability to pick out a signal (true change of a certain magnitude) from the background of natural "noise" variability, and/or 4) decrease our confidence in the magnitude of the signal we will be able to detect), and/or 5) increase cost (see section on monitoring design and representativeness above for more details on random sampling), and 6) often results in clumped samples that are not spatially balanced. Also, considerable seasonal, temporal, or microhabitat-type driven variability is more common for contaminants sampling in water or sediments than has been widely recognized.

However, too often strata or index periods are picked based on untested assumptions about patterns of variability. The more one understands variability in time and space, the better job one can do of making decisions about potential strata and index period widows to sample.

General monitoring design theory holds that "if the response of interest displays substantial variation in one aspect of time or space, but not the other, we need to sample across the variable dimension, but can more or less ignore the other with little loss of information." A typical problem we have in water column measures is that such measures tend to vary across time of year, time of day, and space. So all apparent dimensions appear to vary, and there appear be none that we can automatically ignore. This may present an example where the only way to find a year to year trend would be to narrowly limit index collection time and space-periods in more than one dimension. For example, a protocol might call for sampling only during mid summer low-flow AND only in the morning, and to sample only in full-mixed and/or narrowly-defined microhabitat strata. Again, when making these kinds of decisions, the more we understand about variation in time and space, the better off we are.

7) Revisit and Refine Target Population Details

An earlier section introduced Representativeness and Target Populations. When one is refining protocol details by moving into calculating required sample sizes, it is a good time to revisit target populations. Details on variability in time and space (see

section just above) should be taken into account, and target populations should be identified as narrowly as possible in time and space. Make sure that identified target populations line up with questions, proposed monitoring design, and other factors in this outline. For example, does the monitoring design ensure that the samples will be fully representative of the full range of values in the target population, considering what is know about variability in time and space? Is the target population all bluegill sunfish in the park, or daytime bluegill sunfish between length A and B, only in limited and defined-size-range of small ponds that are road or trail accessible?

8) How Big of a Difference or Change Do We Need to Be Able to Detect?

What magnitude of change (or difference vs. a water quality standard or other comparison-benchmark) do we need to able to detect? Once initial qualitative decisions about desired conditions (DCs) have been made, to intelligently design long term monitoring and to decide and document the magnitude of minimum required-sample-size targets, there is usually no getting around the tough decision of what is the QUANTITATIVE minimum detectable difference (MDD) that we need to be able to detect. Sometimes the MDD is expressed as an "effect size" (hereafter, abbreviated to ES). Whether we call it a MDD or an ES, it is the magnitude of change that we need to be able to detect in order to be able to manage the resource in an optimal, protective, and precautionary manner. This is a decision to be made by NPS staff, not statisticians.

Monitoring Design Sensitivity vs. Measurement Sensitivity

Just as there are detection limits (such as method detection limits/MDLs, see section farther below on measurement sensitivity) for single measurements, on a higher level of organization (multiple measurements), a given monitoring design will have a detection sensitivity (minimum detectable difference, hereafter abbreviated MDD). The MDD magnitude is driven by variability, sample sizes, significance level selected, power magnitude selected, and various other details.

No matter the scale or level of organization, sensitivity always relates to signal to noise ratios and how small of a difference we can detect quantitatively.

Lack of perfect measurement precision contributes variability uncertainty on the measurement scale of concern. True heterogeneity contributes variability on the monitoring design scale of concern. Both contribute to the overall variability of the values recorded.

How would one compare the magnitude of these two contributors to total variability? Measurement precision simply adds variability above and beyond true heterogeneity, so it is already factored in when measurement values are recorded. The question of interest then becomes whether or not measurement uncertainty is so large that it is significantly impacting our estimates of true heterogeneity in variables.

When combining sources of uncertainty in sum of squares equations for combined or expanded uncertainty (N. Taylor and C. E. Kuyatt. 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297, http://physics.nist.gov/Document/tn1297.pdf), one rule of thumb used in the United Kingdom (UK) is that there is no need to add negligible variance [each square of a

standard deviation (SD) is a variance] terms. Although adding them would be the safest (especially if there are many such terms to complicate the issue) in many cases if any of the standard deviations is so small that their contribution to overall uncertainty is negligible (the standard deviation is at least five times smaller the standard deviation of the next largest contributor to uncertainty), they may often be ignored (hereafter the 1/5th rule, from a UK publication that discusses NIST/ISO uncertainty, United Kingdom Accreditation Service, 2000. The Expression of Uncertainty in Testing, UKAS Publication ref: LAB 12, http://www.ukas.com/Library/downloads/publications/LAB12.PDF).

If there are more than two major contributors to variability, do the sum of squares with all variance (SD squared) terms included to see if including the relatively low magnitude SDs changes the results appreciably.

Beyond the issue of having more than two major contributors to variability, there are other issues that might make the 1/5th rule not work optimally in every situation. For example, in some types of reconnaissance-level sampling, it is impossible to collect enough samples to accurately define natural variability (either temporal or spatial) within the time frame and funding available. On the measurement scale of concern, unless one is looking over a long time period and multiple QC samples, one often does not have a large enough sample size to accurately estimate a standard deviation for measurement precision. In fact, at first one typically just has sample size 2 (difference is expressed as a relative percent difference---RPD). Why is this important? When either the enumerator or the denominator (and especially both) are not good estimates of the SD, one can't accurately judge the 1/5th threshold. The 1/5th rule depends on good estimations of the standard deviations. Standard deviations are typically not well estimated at small sample sizes (below 25-30 and especially below 7-10) or when the values used to generate the SD for true environmental heterogeneity do not represent the full range of conditions of the representative target population being sampled.

If sample sizes are too low or if calculated SDs are not representative of the target population, these faults should be corrected before putting too much weight on the results of the 1/5th rule of thumb.

This factor of 5 (expressed as a SD) is generally consistent with other signal to noise rules of thumb. Most of these rules of thumb state that (for accurate measurement) a signal should typically be 3 to 10 times greater in magnitude than noise (see Part B for more detail and several examples). Statements such as "errors in the analytical measurements should be no greater than the natural variability of the parameters of interest" should be rejected since such errors should be at least 5 times lower (when expressed as SDs).

What about other summary statistics used for variability? Can one also see if either a coefficient of variation (CV) or a relative standard deviation (RSD = CV*100) is 5 times lower than their counterparts, when estimating true environmental variability (on measurements of different samples)? No, these are different. The 1/5th rule of thumb should be used with SDs only. Using this rule for other summary statistics produces different results.

Calculate Monitoring Design Sensitivity

Why determine detectable differences? Generic VS guidance has suggested that networks "List the specific, measurable objectives for each vital sign selected for monitoring, and wherever possible, give the threshold value or "trigger point" at which some action will be taken" (Outline for Vital Signs Monitoring Plans, 2004, http://science.nature.nps.gov/im/monitor/docs/monplan.doc). To be precautionary in preventing major impacts or even a resource collapse, monitoring networks typically have to be able to detect a change smaller than the chosen thresholds or trigger values (see more detail below).

Choosing minimum detectable differences (MDDs) on the monitoring design scale of concern is usually done in an iterative manner. First choose an initial MDD that seems reasonable (maybe pick 20-40% as a change in a mean as a starting point if nothing else seems more logical). Next, have the network quantitative ecologist run the numbers (step 12) based on other decisions for steps 7-11 to see if the monitoring design will be able to detect a change that small. Often initial decisions on sample sizes, sample numbers, sample locations, and detectable differences simply will not work and adjustments in one or more of these factors (and/or in alpha or beta) are be needed to improve detection probabilities.

A minimum detectable difference (MDD) is typically a minimum detectable difference between means (or medians, or the summary statistic chosen and a water quality standard or other benchmark), all in original units of measure.

The effect size (ES) is similar to the MDD but is expressed in standard deviation units rather than original units of measure. It is usually the minimum detectable difference between means (or between a mean and a water quality standard or other benchmark) divided by best estimate of the true (but unknown) population standard deviation. That best estimate is most often a pooled standard deviation that covers a broad range of conditions. The pooled standard deviation (SD) for two or more samples is essentially the square root of the average variance; see full equation for a pooled SD at http://science.widener.edu/svb/stats/descript.html. The ES result is the magnitude of change expressed as the number of standard deviations. One then multiplies the result x 100 to get the result expressed as a percent of the standard deviation. So the final ES is in % units rather than original units of measure.

In psychology a "large" standardized effect size of 80% in SD units. This would often be a small change in field biology scenarios. One ES advantage is that comparisons of effect sizes for different measures or vital signs can easily be made between different vital sign measures and time frames, since all ESs are expressed in SD units. Also, no initial estimates of variability are needed. This is a big advantage if one is making initial calculations before any credible estimates of variability are available (Cohen, J. 1988. Statistical power analysis for the behavioral sciences. Lawrence-Erlbaum, Hillsdale, N.J.).

However, since environmental variables are often not normally distributed and sample sizes are often small, such estimates are usually used only for very rough initial estimates, for comparing effect sizes between indicators with different original units of measure, and for looking at effect magnitudes from different angles (not using original units).

A major drawback of using ESs in SD units is that if one chooses a certain ES magnitude (say 80% for example), one will choose the same sample size regardless of

the accuracy or reliability of the measuring instrument or the true variability of what is being measured. This is clearly not ideal, and one more reason to also look a power or detectability in original units as soon as possible. Lenth also points out that one should use power prospectively, put science before statistics, and do pilot studies (Lenth, R. V. 2006. Java Applets for Power and Sample Size [Computer software]. Last opened February, *2006,* from http://www.stat.uiowa.edu/~rlenth/Power).

There are a few isolated cases where the NPS has logical reasons for the need to be able to detect very small change, such as a 5% MDD as change in means in original units (not SDs). For example, for air quality goals, a 5% change in visibility was shown in human studies to be "perceptible." The Clean Air Act states that "visibility impairment" is defined as "any humanly perceptible change in visibility from that which would have existed under natural conditions." [Federal Land Managers Air Quality Related Values Work Group Document, section D2 (visibility), http://www2.nature.nps.gov/air/Permits/flag/docs/FlagFinal.pdf].

However, usually it is difficult to detect changes that small in means or original units, so it is more common to try to detect 20-40% differences.

In deciding how big of a change or difference monitoring needs to be able to detect, it is helpful to consider the type of park(s) being monitored and park-specific management goals. In the variable worlds of aquatic biology and water quality, detecting a 5% change, either as a MDD or ES, would usually be impossible or require so many samples that it would be prohibitively expensive. Keep in mind that the smaller the MDD or ES one is trying to detect, usually the more samples one would have to take (more costly) and the more difficult it is to find a strata where the variability is low enough to allow detecting such a small change. Therefore, it is usually not advisable or often not possible to detect extremely small changes (1-5%). Consider the following more typical scenarios:

> Scenario 1: The resource to be protected is an endangered species or a very highly prized and rare resource in a relatively pristine area of a park having natural resource protection as a key goal: In this case, a park resource stewardship plan might logically call for a relatively high degree of protection and management precaution. Such a park might even choose to protect such a resource very stringently. Anything above a negligible or even a detectable impact of concern might not be considered acceptable relevant to desired conditions (DCs). The minimum detectable difference (MDD) the park might want to detect might be a relatively stringent 5-25% change in means in original units of measure. In SD units, the ES size the NPS might want to detect might be as small as a 10-30% change when the change-magnitude units are the number of standard deviations expressed as a %.
>
> Scenario 2: The resource to be protected is a population of an aquatic species at a typical national park, but the species is very common in the region and/or nation (for example, a bluegill sunfish). In this case, the park might designate a less-stringent MDD, such as a more common 20-50% change in means, or an ES change-magnitude of 30-70%, when the units are a percentage of the magnitude of the standard deviation.

> Scenario 3: The natural resource to be protected is general water quality or a population of a common aquatic species at a historical park in a highly urbanized or highly farmed area. Here the water quality and aquatic habitat is such that there is little or any hope of ever reaching "unaffected by modern civilization" status, and this is reflected by state water quality standards and biocriteria that are less stringent than one would find in less-impacted areas of the country. Perhaps the biota to be protected is short-lived and highly variable even in pristine areas. In this scenario, the park might decide that only larger changes can or need be detected. The park might therefore want to be able to detect a MDD of a 40-80% change in means, or an ES change of 70%-90% when the units are a percentage of the magnitude of the standard deviation.

The examples above are mentioned only to give monitoring planners a very rough idea of some typical ballpark (starting-point) ranges of values. Whenever one has first developed a logical and defendable park-specific MDD or ES, of course those values should be documented and used instead of the examples above.

What if the park simply has no idea whatsoever how big of a MDD they need to be able to detect to protect important resources? Try detecting changes or differences of the magnitudes mentioned above as a starting point.

For some indicators, the ability to detect a 20% change in means in one year or even multiple years might require too many samples and exceed the budget. A network could consider the approach that Channel Islands National Park adopted as a starting point for VS monitoring. That park adopted a preliminary goal of being able to detect a 40% change in means from year to year, with alpha (Type I error, the polluter's risk) of 0.05 and beta 0.2 (power 80%, the Type II error, the conservationist's risk), with the stated idea that the values could be changed later if needed, as lessons were learned (Gary Davis, NPS, Personal Communication, 2005, based on Paper to be put on CHIS website).

In common situations where one is not trying to protect an endangered species or something else especially rare or valuable, it is uncommon to try to detect a biological effect size smaller than a 20% change in means. That default is often adopted as a "de minimis" (the law cares not for little things) starting point when one cannot logically come up with something better. A 1992 paper suggested that it was difficult to find cases where a state or federal regulatory agency had prosecuted anyone for a biological effect size of less than 20% of the mean on non-human or non-endangered species. This was true regardless of whether the population, community, or ecosystem level was being considered (Suter, G.W. II, A. Redfearn, R.K. White and R.A. Shaw. 1992. Approach and strategy for performing ecological risk assessments for the Department of Energy Oak Ridge Field Office Environmental Restoration Program. Martin Marietta Environmental Restoration Program Publication ES/ER/TM-33, Environmental Restoration Division Document Management Center Environmental Report (ER), Environmental Sciences Division (ESD) Publication 3906, Oak Ridge National Laboratory, Oak Ridge, TN, pp. 8-9).

Funding limitations should not be an excuse to monitor a large number of sites (or a large number of measures) poorly. It would be better to monitor fewer sites and fewer vital signs well). Or, as said in more technical terms in the 2005 SWAN phase III report

(op.cit), “It is better to gather sufficient data on a smaller area of inference than inadequate data on a larger scale of inference” and “It is better to gather data of sufficient quality on fewer vital signs than insufficient data on many of them”

Again, in water quality, one can sometimes reduce variability by carefully picking integrator variables (for example, benthic macroinvertebrates) and by narrowly defining target populations (for example the populations in riffles only, or on snags only), in narrow index time-periods only). This tends to reduce variability compared to randomly sampling water column variables in all habitats at all times of year.

Reducing variability with narrow definitions of strata, target populations, and extent of inference, can helps one achieve more uniform magnitude of variability between-years and thus helps one pass the straight-face (common sense) test when claiming that the results from samples from multiple years can be lumped before estimating a proportion, and still represent a valid single-sample. In done right in context, this can facilitates the justification of rotating panel designs that can reduce the number of samples needed per year. Reducing variability also helps one detect changes of a size of concern.

Another strategy is to move a continuous monitoring sonde around on a rotating panel basis to get required sample sizes and to document temporal variability. In concert with oft-repeated generic NPS vital signs guidance, networks should design monitoring that produce credible information with available funding, but identify areas for expansion. As we deliver value to parks and increase our partnerships, additional resources may be available to support a larger program.

No matter how the quantitative levels are developed, once the MDD or minimum detectable ES is developed as our best quantitative estimate of the change that the park or network would like to be able to detect, a reality check comparison should be made between that value, the current condition, and the DC. What is the amount of change it would take to get to different levels of compliance with water quality standards or to get to levels that would cause a resource collapse (if known)? All such goals should logically reflect park management goals. In the park planning process, such a comparison might be made in the newer-style NPS resource stewardship plans.

If the initial results seem unsatisfactory, monitoring networks can sometimes use a more precautionary by lowering beta to 0.05 or 0.1, or by performing inequivalence testing rather than standard null hypothesis testing. See details further below. In any case, the starting point MDD can be changed to a different % based on park protection goals or ecological and other lessons in an iterative (adaptive management) fashion.

9) What Initial Statistics Will Be Used?

Do the questions to be answered call for detecting A) a difference between two means, B) a difference between a mean and a water quality standard, C) a trend (over how many years), or D) an estimation such as a confidence interval (CI) about a mean or a proportion? Will parametric or nonparametric statistics be used? The answer to these questions will help drive which statistics will be used and how sample sizes are calculated. For more details, see section IV-C of Part B.

10) Choose Desired Detection Probability/Statistical Power and Degree of Confidence

What is the degree of confidence we want to have in detecting our chosen change magnitude of concern? This is a decision for park and other NPS staff to make with the help of the network quantitative ecologist. Do we want to have 80%, 90%, or 95% confidence that we can detect the change magnitude of concern? The decision may vary with rareness and special value of the resource being protected, how pristine the area is, or the variability of the vital sign or measure. For a measure that is highly variable even in a variability-reducing stratum, it may difficult or impossible to detect a 20% change in means or medians with 90% confidence.

Typically the network would work with a NPS quantitative ecologist (and later with a statistician) in an iterative manner to decide workable probability of detection percentages. In other words, once an initial decision has been made about desired probability of detecting the effect size of concern, it is typically necessary to run the numbers to determine the actual probability of detecting an effect size of concern given the monitoring design, the variability of the parameters, the statistics chosen, and other factors discussed herein. If the detection probabilities turn out to be unacceptably low, changes in the overall monitoring design, analytes to be measured, sample sizes, sample locations and strata, sample timing, etc. may have to be made. It is common to have to repeat sample size calculations several times before satisfactory compromises between power, budget, and choice of effect sizes can be made.

If hypothesis testing is not done, but estimation (say of a confidence interval about a mean or proportion, see step 12 below) is done instead, the degree of confidence is typically expressed by the magnitude of the confidence interval rather than beta.

11) Choose Significance Level (alpha)

What is the desired probability (1-significance level) to avoid falsely detecting a change or difference of the magnitude of the MDD or ES (type I error, the polluter's risk)? Too often what looks like science---choosing an error rate for a statistical test for water quality assessments---is an unrecognized public policy decision (Shabman, L. and E. Smith. 2003, Implications of Applying Statistically Based Procedures for Water Quality Assessment." *Journal of Water Resources Planning and Management*, 29(4): pp. 330-336, see rest of quote and related discussion on page 176 of McBride, G.B. 2005. Using Statistical Methods for Water Quality Management: Issues, Options and Solutions. Wiley, NY, 313 pp., http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471470163.html).

In the p-value culture of many journals in the past, significance level (alpha) has traditionally been set at 0.05. In the p-value culture of many journals in the past, significance level (alpha) has traditionally been set at 0.05. Although recently there have been many articles explaining why automatically doing this (and over reliance on p values in general) is not a good idea, the culture persists. For more information on the problems, see Part B and some of the many recent Internet discussions, such as http://www.npwrc.usgs.gov/resource/1999/statsig/stathyp.htm, http://www.niwa.co.nz/rc/prog/stats/issues.pdf,

http://her.oxfordjournals.org/cgi/content/full/14/6/713, and
http://www.warnercnr.colostate.edu/~anderson/null.html.

When doing inequivalence testing, 0.05 is fine. For other types of testing, the NPS is typically even more concerned with false negatives (wrongly concluding no impact or no change when a change or impact has, in fact, happened) than false positives (wrongly concluding impact or change when a change or impact has not happened). Therefore, network quantitative ecologists may at times specify small beta levels, even if they are smaller than alpha (see Part B for more details). Unless one was using an inequivalence test, instead of blindly insisting that alpha always be 0.05 according to tradition, monitoring planners might choose 0.1 or 0.2 for alpha. Beta might then be 0.05 or even 0.01. As a resource conservation agency, the NPS should not be less worried about beta (the conservationist's risk) than the alpha (the polluter's risk).

12) Use Simple Calculators to Make Initial Estimates of Required Sample Sizes

Due to recent breakthroughs, it is now easier for quantitative ecologists to use sample size equations and calculators on the Internet to get a rough idea of sample sizes needed. Power should not be ignored, so be sure that any MDD calculator used has inputs not only for alpha, but also for beta (1-power) and for the standard deviation (SD) or variance). If ES calculators are used as a first step, be sure the calculator has inputs for alpha AND for beta (1-power) and follow up with MDD calculators to look from a better angle when a SD is available.

One typically does sample size calculations in an iterative manner, trying various samples sizes until the answer stabilizes (rounds up the same whole number) and/or by playing what-if games. If one already has a starting sample size, the McBride probability of detection calculator (op. cit.) can also be used to estimate the effect size that can be detected with various sample sizes, probabilities, stated significance levels, and various types of tests. Start by clicking on the effect size button and then choose either one or two groups.

> CAUTION: There are many sample size and power calculators on the Internet. Before using them, we suggest checking them against Zar's examples to make sure you can get the right answers. Some of the Internet calculators appear to use the wrong equations and/or give the wrong answers. One used the middle t-value instead of the (correct) upper (one-sided) t-value for power. Also, it is easy to misunderstand some of the input variables or their format, which is why detailed step-by-steps are given in Part B. As will be repeated for emphasis, hypothesis test sample size estimators that don't take into account power (1-beta, or analogous probability of detection for inequivalence tests) rates should not be used unless otherwise justified.

Step-by-step examples on how to use some of the more user friendly Internet calculators to get the same answers as the Zar examples are summarized below and in more detail in section IV-C of Part B.

Simple sample size calculators provide just rough (but far better than nothing) starting point estimates of needed sample sizes. Some would argue that the same is true for more complex simulation approaches.

Perform Different Initial Simple Calculations Depending on the Scenario:

We recommend estimating sample sizes needed with a multi-step approach. No matter how one is estimating required sample sizes, it is important to keep documenting and checking assumptions at each step along the way. Calculations for sample size requirements, like all inferential techniques, are based on certain assumptions. Be sure to discuss assumptions in eventual discussions with an applied statistician (see step 16).

The first few steps listed below relate to hypothesis testing, since these tests are still commonly used and since many are familiar with them and with the Zar equations (details below). This does not imply that standard null hypothesis testing should be a first choice for analysis. In fact, the standard null hypothesis test (especially when power is not controlled) is usually not an optimal choice for ecological field work (see Part B for details), and even when the inequivalence test or other good options involving controlling power are chosen, test results are often not definitive by themselves and typically should be used only as one of many lines of evidence considered in ecological resource management decisions.

Nevertheless, parks often want to compare results from one site to another or from one time period (say one 5 year period of drought) to another time period (for example 5 relatively wet years). Sometimes parks even want to compare one year to the very next. More often, if they visit plots two years in a row and then rest them for several years in a rotating sampling approach, they might wish to compare the average of these two year periods to the next two year sampling period. In several of these scenarios, the first few items below are relevant for estimating needed sample sizes.

Before using calculators, it is very important make sure you understand exactly how to input variables need to be formatted and entered and check some example problems where the correct answer is known before proceeding.

It is also a good idea to perform multiple calculations from the options listed below to “look at the issue from multiple angles” and to see if they are all close. If they are not, an input variable may be wrongly formatted. If they are close, take the highest sample size estimate to be precautionary, in the absence of a better rationale.

Sample Size Calculations for Nonparametric Procedures

Most of the sample size (and related power) calculators use the t-distribution and assume normality (of the sampling distribution of the means, *not* of the distribution of values themselves). What if assumptions are not met and nonparametric tests will be used? Nonparametric sample size and power estimators are not easily available.

Most experts have suggested that the power and sample size requirements for nonparametric tests are usually not all that different than those for parametric tests. Some suggest that one could use the parametric calculators for first estimates of sample size and then add a small amount (perhaps 5%) to be precautionary if one is unsure of the normality of the distribution being sampled. The initial sample size estimates are usually

rough anyway, and if one wanted really exact values, one might go to more complex methods (such as simulation) to estimate sample sizes.

Others see adding a small amount to the sample sizes as unnecessary, since field environmental data is usually badly skewed. They point out that when the assumptions of the t-test are badly violated, the Mann-Whitney test has **more** power than those tests that require assumptions of normality (Zar, op.cit). The Wilcoxon Sign Ranks test for two independent samples has about a 95% efficiency compared to a t-test when the distribution is normal.

The key point is that field collected data never have exactly normal data. If approximately normal, the two tests have equivalent power (small departures from normality offset the 5% difference in power), and so require the same numbers of samples. Therefore, there is typically little need to increase sample sizes by even 5% before performing nonparametric statistics. When one transforms to the log scale, one can usually still use the parametric calculators (just don't back transform means). If one was using the log transformation, how many samples would be required to see a 10% difference in medians (geometric means) in the original units? One way to estimate this is as follows: 1) Since differences in mean logs corresponds to a ratio of geometric means in original units, first estimate the geometric mean of the original data. 2) Next, determine the number of samples needed for a 10% difference in medians by using the parametric calculators. 3) Determine the difference that needs to be detected on the log scale as the difference between the geometric mean (median) and 1.1 times the geometric mean. Percent differences (in original units) are additive differences in log units. So the traditional sample size formula can be used in log units to determine how many samples are needed to see a 10% difference in medians (not means) in original units. (Dennis Helsel, USGS, Personal Communication, 2006). See Part B for more additional documentation and detailed information for different multipliers for different nonparametric tests in the unlikely case that one runs across normally distributed distributions in environmental sampling.

Many investigators use the t-distribution calculators for data that is log-normally distributed and then log transform environmental values before using parametric tests, especially when sample size is above 20-30. However, this should not be an automatic choice. Sometimes nonparametric analyses are a better choice, and simulation can sometimes do a better job of estimating needed sample sizes. Also, be careful to avoid back transformation bias. If the mean is the main focus and the test is to be done on transformed values, then run the sample size estimators with transformed values, and don't report back transformed means, standard deviations, minimum detectable differences MDDs in means, or variances. MDDs in geometric means (medians) can be used.

> Note: Again, the key point is that back transforming a log mean gives the geometric mean, which estimates the median (not the mean!) of the original units assuming the logs are symmetric. If logs are nearly symmetric, sample size calculations in log units will be a good approximation to true nonparametric sample size estimators (Dennis Helsel, USGS, Personal Communication, 2006).

The Helsel and Hirsch text book (Helsel, D.R. and R.M. Hirsch. 2002. Statistical Methods in Water Resources. US Geological Survey Techniques of Water Resources Investigations, Book 4, Chapter A3, available for free on the Internet at http://water.usgs.gov/pubs/twri/twri4a3/pdf/twri4a3.pdf) does not go into sample size calculations in detail. However, chapter 4 provides good detail on using nonparametric hypothesis tests, even with small sample sizes.

Before a Good Estimate of the Standard Deviation is Obtained

If a hypothesis test is to be used, perhaps as one line of evidence, we suggest starting sample size estimation with the McBride detection calculator (http://www.niwa.co.nz/services/statistical/detection). The McBride calculator is a good first step since: 1) it allows one to make initial estimates of sample sizes if there is no (or no good) estimate of variability, and 2) it allows one to look at probabilities of detection of various effect sizes in not only t tests but also in the generally precautionary and more optimal inequivalence tests. Impacts may change not only means but also SDs (more impact often leads to more variation). Therefore, it is not a bad idea to pay close attention to changes in standard deviations (SDs) and in changes expressed as a percent of a SD. In the McBride calculator, remember to express effect size as a percent of the standard deviation in the calculator input. So, if ES = difference/SD = 0.69, express the effect size as 69%. Be sure the SD used is a best estimate of true heterogeneity of different samples, not generated from repeat sampling of a single sample.

After a Good Estimate of the Standard Deviation is Obtained

Once one has a decent estimate of a SD, do another quick calculation based on Zar's equations (Zar, J. H. 1999. Biostatistical Analysis. Prentice Hall, Upper Saddle River, New Jersey, USA). UCLA Internet calculators can be used to roughly determine sample sizes for t tests, given known variability. These power and sample size calculators offer the advantage of having effect sizes in original measurement units as input variables.

Computing sample size needed to detect a stated difference (minimum detectable difference = MDD) between a mean and a water quality standard or other benchmark: Use Zar equation 7.8 and one-sample, normal-distribution, two-sided, sample size calculators after checking their accuracy vs. Zar example 7.7 (page 107, Zar 1999, op. cit.). When checked in July 05, the UCLA 1-sample, normal-distribution, two-sided, sample size calculator (at http://calculators.stat.ucla.edu/powercalc/normal/n-1/) produced the same answer (19) as Zar gives in example 7.7. Input data like this: To calculate sample sizes, put a question (?) mark the box for sample size, input alpha (example 0.05) and beta (example 0.9 for 90% power, make the minimum detectable difference the difference the mean (in original units, enter 100 and 101 for a difference of 1.0 gram in the example), and then choose two-sides.

Note: If one already has a proposed sample size, one can also use's Zar's (rearranged) equation to solve for MDDs: Zar's minimum detectable difference (equation 7.9, Zar op. cit.) for one sample vs. a water quality standard as follows: MDD = square root of [sample variance/n] * (middle-probability two-tailed t-value for 1-alpha given-1 degrees of freedom and probability chosen + upper, one-tailed t-value for 1-beta given n-1 degrees of freedom and probability chosen)^2. Again, the UCLA calculator at http://calculators.stat.ucla.edu/powercalc/normal/n-1/ appeared to give the same answer as Example 7.8, Zar (op.cit, page 107), but be careful how the inputs are made to get it to work right (see Part B).

Computing the sample size needed to detect a pre-determined magnitude of difference between two means (minimum detectable difference = MDD): Use Zar equation 8.22. Planners may be able to use an Internet calculator after confirming they give the same answer as Zar provided (Zar 1999, op. cit.). When checked in 2005, the UCLA two-sample normal distribution, equal variances, sample size calculator at http://calculators.stat.ucla.edu/powercalc/normal/n-2-equal/ gave the same answer (45 for each sample) as Zar gives in example 8.4 (page 134, Zar 1999, op. cit.) for equation 8.22. To calculate sample sizes, put question marks in both boxes for sample sizes.

If Variances Are Unequal

It would not be unusual for a standard deviation (or variance) to be different in one time period versus a later time period. If one has unequal variances one can use the UCLA calculator at http://calculators.stat.ucla.edu/powercalc/normal/n-2-unequal/. Other internet calculators also cover unequal variances, including

An Iowa State calculator (http://www.stat.uiowa.edu/~rlenth/Power/index.html**).**

A University of Wyoming calculator by Dr. Kenneth Gerow, Professor of Statistics (http://www.statsalive.com/).

This last calculator is a free MS-Excel macro that includes options for the following scenarios: 1) paired samples, 2) either SD or variance proportional to the mean, 3) for different sample sizes, and 4) either equal or unequal variances. This more advanced calculator includes instructive example plots in help screens. Be sure to read the help screens carefully as the input variable choices are not quite as quickly understood as some of the other calculators mentioned above. An advantage of Gerow's calculator is the extra options. For example, SDs are often proportional to the mean (SDs often go down as sample sizes go up) and means also tend vary by sample size. The Gerow calculator provides a way to easily draw graphs on the SD/mean and variance/mean comparisons. Planners often decide to do paired samples (which gives more power but fewer degrees of freedom), and sample sizes are often not equal. With careful input, we were again able to use this Excel calculator to get the same answer (45

for each sample) as Zar gives in example 8.4 (page 134, Zar 1999, op. cit.) for equation 8.22 for an equal variances and equal sample size problem.

Using paired samples does not insulate one from problems related to small sample sizes or asymmetric distributions. "Power decreases as the variance increases, decreases as the significance level is decreased (i.e., as the test is made more stringent), and increases as the sample size increases. A very small sample from a population of paired differences with a mean very different from 0 may not result in a significant t test statistic unless the variance of the paired differences is small. If a statistical significance test with small sample sizes produces a surprisingly non-significant P value, then a lack of power may be the reason. The best time to avoid such problems is in the design stage of an experiment, when appropriate minimum sample sizes can be determined, perhaps in consultation with a statistician, before data collection begins" (Northwestern University Website at http://www.basic.northwestern.edu/statguidefiles/ttest_paired_ass_viol.html).

Again, it is important to consult a professional statistician before finalizing sampling designs. Although the calculators listed above would often be fine as a first cut (illustrating which variables or strata that are just way too variable to allow networks to detect a difference of concern), all such simplified calculators should be a used mostly as a first step. As more data is collected and monitoring plans began to be refined, networks should consult with a professional APPLIED statistician for more advanced fine tuning of samples size and power estimations. Those with more advanced expertise may choose to use more advanced methods (such as the Gerow calculator just above or simulation approaches).

Composite Samples, A Special Case

How does one determine statistical power in relationship to sample sizes when compositing many individual fish into single composite tissue samples for contaminants analyses? In 2000, EPA provided look-up statistical power tables that illustrate that as "a factor similar (sic) to a coefficient of variation (CV)…as the ratio of the estimated population standard deviation to a screening value (SV) increases (i.e. SD/SV), the statistical power decreases" (see http://www.epa.gov/ost/fishadvice/volume1/v1ch6.pdf).

Gilbert has a whole chapter on compositing, providing complex formulas for estimating the variance of means and required sample sizes (Gilbert, R.O. 1987. Statistical methods for environmental pollution monitoring. Van Nostrand Reinhold Co., pages 6 and 72).

In considering composite samples and sub-samples, making sense involves understanding what Cochran calls "relative precision," the ratio of the variance from the combined sample (local small area plus larger area) over the variance from the larger area sample (W.G. Cochran. 1977. Sampling Techniques, 3rd edition, John Wiley & Sons, New York).

In composite sampling, it is also helpful to understand differences in means and variances (between compositing and not compositing) on a few pilot-scale (trial) samples before finalizing composite schemes for very large and expensive monitoring projects. For more details, see Part B.

When In Doubt, Throw It Out:

Consider throwing out analytes or measures where the variability in pristine sites is so high (even using strata or response design details that reduce variability the most) that one would never find a trend or difference of biological concern given funding limitations. For the same reason, consider throwing out variables having unacceptable levels of measurement uncertainty. In other words, consider using only measures having acceptable levels of measurement precision, acceptably low detection limits, and acceptably low measurement bias (see chapters on those topics further below). This theme is so important that we will revisit in future steps.

Inequivalence Calculators

One can use the McBride detection probability calculator to compare two-sample inequivalence, equivalence, or standard null hypothesis t tests (http://www.niwa.co.nz/services/statistical/detection) to estimate the probability of detection of various effect-sizes (expressed as a % of the true standard deviation) given the sample size chosen. Play what-if games with the precautionary inequivalence option vs. other options given various effect sizes and sample sizes to see how detection probabilities change. In a perfectly normal population a standard deviation might be as much as 3 times smaller than a mean, so an effect size of 50% of the standard deviation, considered a moderate effect size in high sample size psychological studies, might be considered small when compared to the mean in smaller and skewed data sets more typical of water quality or aquatic ecology, unless one was near a resource collapse threshold. In typical (skewed) field environmental data sets, a standard deviation can be half or even equal to the magnitude of the mean, so after using the McBride calculator, translate the values back to percentages of the mean or median to get a reality-check look at the effect size from a different angle. The equation used to translate back to original measurement units is effect size = the difference in means or the difference between a mean and a comparison water quality standard in original units of measure = the effect size (as a percent of the standard deviation) times the standard deviation, all divided by 100.

Proportions

Calculating proportions well is notoriously difficult and usually should not be attempted with a sample size less than 25-50. Calculating the exact needed sample sizes for the estimation of a proportion is typically done in EMAP style surveys. The proportion of stream length or miles impaired is an objective of interest to some parks and monitoring networks. Part of the appeal is that it can relate to GPRA and other more general goals (desired conditions, condition assessment percentiles, ecological thresholds, etc.). One typically should use probabilistic monitoring designs for questions such as: "What percent of stream miles are impaired?"

One can make initial estimates of needed sample sizes with table 1 and the equation at http://www.epa.gov/nheerl/arm/surdesignfaqs.htm#manysamples. In this case, sample size calculations depend only on the proportion and desired % confidence (a z-

distribution confidence interval on the proportion) required. Note that EMAP is using the word "precision" in a nonstandard (vs. QA/QC or NIST/ISO terminology) way here. What EMAP means by precision is a z-distribution confidence interval surrounding a summary statistic (a proportion in this case), rather than measurement precision. The *t*-distribution that is used for smaller sample sizes for means can't be used for proportions because it is not applicable to sampling from a binomial distribution (the same is not true for the z-distribution at larger sample sizes). The confidence interval equation should only be used at sample size 25 or above (n = 50 is the recommended default) but does not depend on a normal distribution. For more details, including a step-by-step example for using the equation, see Part B. A minimum sample size of 25 relates to different sites rather than to replicate times of day at the same site. However, same-site temporal variability will have considered when making data analysis decisions, to help make sense of the data related to standards exceedances. A key common sense question is how much change in time or conditions can occur before the sample is no longer one sample (and maybe sample size is no longer big enough to estimate a proportion reasonably well).

Trends

Estimating sample sizes needed for trends can be tricky, and hints of some kind of trend can almost always be found in long term monitoring. Again, even though statistics used for trends are often nonparametric tests (like the seasonal Kendall Test), sample sizes need are often first approximated with parametric calculators.

Although nonparametric tests are favored by many in USGS, other investigators (and some software, including that used by EarthSoft and Lakewatch, no government endorsement implied) have used parametric regression techniques distinct from simple linear regression in that the data is deseasonalised (G. Barnes, 2002, Water Quality Trends in Selected Shallow Lakes in the Waikato Region, 1995. Environment Waikato Technical Report 2002/11, http://www.ew.govt.nz/publications/technicalreports/documents/tr02-11.pdf). The same method is one of the methods used in Lakewatch programs in various parts of the US, including Florida (http://lakewatch.ifas.ufl.edu/2003DataReport/FLWVol2_IntroIndex.pdf). To be conservative, Florida Lakewatch looks at trends using multiple angles and tests and only calls trends if multiple methods indicated a trend. Lakewatch software, including a free 30 day trial, is available at www.lakewatch.net (no endorsement implied, we have not tried it).

The deseasonalised parametric tests (paragraph above) are mostly based on the work of Noel Burns, who has clarified that he does not emphasize power but instead looks at multiple lines of evidence (chlorophyll, Secchi Disk, TP, TN) to see if most point in the same direction (towards or away from eutrophication). Burns statistically evaluates the coherence of the trends in the 4 variables and translates this coherence, or lack of it, into the probability of change in a lake. Burns also emphasizes looking at the data from different angles (original data and various statistics or transformations) and particularly the seasonal pattern, which can be quite different for each variable, and for different lakes. Burns statistically evaluates the coherence of the trends in the 4 variables and

translates this coherence, or lack of it, into the probability of change in a lake (Noel Burns, EarthSoft Consultant, Personal Communication, 2005).

Although I have not seen direct comparisons done, I would expect the deseasonalised method above to produce reasonably similar results to the seasonal Kendall Test (Graham McBride, NIWA, New Zealand, Personal Communication, 2006).

Both two-sided (trends either way) and one-sided (trend in one direction only) sample size calculators are available from EPA (beta version) at http://www.epa.gov/earth1r6/6wq/ecopro/watershd/monitrng/qappsprt/sampling.htm.). Parks are typically interested in trends whether the trend is going up or down, and these call for using two-sided equations.

> **Two-Sided Trend Applications:** The equation used by EPA and others to estimate adequacy of sample sizes for trends is n = 12 * (sample variance of the de-trended series) * $[t_{a2(n-2)} + t_{b(n-2)}]^2$/trend magnitude^2, where $t_{a2(n-2)}$ is the two-tailed middle-probability t critical value for sample size n-2 and where $tb_{(n-2)}$ is the one-tailed upper t critical value for sample size n-2 using alpha of 0.05 and beta of 0.1. De-trending techniques in Excel are explained at http://www.bized.ac.uk/timeweb/crunching/crunch_analysis_illus.htm. One Australian state will not call a trend unless 1) the sample size was adequate (according to the preceding equation) and 2) the result of a Kendall test for trends indicates a trend. These issues and others (autocorrelation, etc, are discussed in a plain-language Internet document (Western Australia Water and Rivers Commission. 2004. Statewide Assessment of River Water Quality Methods, http://apostle.environment.wa.gov.au/idelve/srwqa/methodology.htm)**.**
>
> **One-Sided Trend Applications:** The one-sided equation (input variables bolded) is very similar but uses a one-tailed t value: n = 12 * (**variance estimate**) * [ta,v) + tb(1),v)]^2, all divided by the trend magnitude squared, where ta,v is the one-tailed upper probability t critical value for sample size n-2 and where tb(n-2) is the one-tailed upper t critical value for sample size n-2. As can be seen in the equations, in the one-tailed case (t alpha) has a right-tail area of alpha. In the two-sided test, t alpha has a right-tail area of alpha/2. Regardless of whether the one or two sided choice is chosen, in the EPA calculator (op.cit.), **significance** (example alpha = 0.05) and **power/detection probability** (example 0.9) may be changed in the fields just below the equation, which automatically changes both the alpha and beta terms accordingly. The one-tailed equations are of special interest to EPA for regulatory questions, such as "Did best management practice implementation improve historically poor water quality in a watershed by some given percentage?" or "Did a new industrial discharge result in declining water quality?" For long term Vital Signs monitoring, two-sided trend tests are usually more applicable to project goals.

Statistical issues for trends in water quality work, including Kendall and seasonal Kendall tests, Sen's Slope Estimator, and other options for assessing trends and autocorrelation are discussed in more detail in the following references:

1) Graham McBride. 2005. Using Statistical Methods for Water Quality Management: Issues, Options and Solutions. Wiley, NY, 313 pp., http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471470163.html,).

2) Helsel, D.R. and R.M. Hirsch. 2002. Statistical Methods in Water Resources. US Geological Survey Techniques of Water Resources Investigations, Book 4, Chapter A3. This textbook is available for free on the Internet at http://water.usgs.gov/pubs/twri/twri4a3/pdf/twri4a3.pdf. Formerly published (now out of print) as Helsel, D.R. and R.M. Hirsch. 1992. Statistical Methods in Water Resources. Studies in Environmental Science 49, Elsevier Publishing, NY).

3) EPA. 2000. Guidance for Data Quality Assessment Practical Methods for Data Analysis, EPA QA/G-9, EPA No. EPA/600/R-96/084, http://www.epa.gov/quality/qs-docs/g9-final.pdf, and Zar, 1999 (op. cit.).

Rethink Detectable Difference Goals for Trends

When considering trend analysis options, rethink how big of trend needs to be detected. EMAP tries to detect a 20 % minimum detectable difference in means over 10 years. They wanted to be able to detect a 2% a year change over 10 years. In another EPA similar EPA example, one criterion for picking an indicator was whether or not it could detect a 20% change in ecological condition over a 10-year period with 90% confidence (Kurtz et al. 2001, op. cit., citation above in objectives section).

For highly valued or rare species not characterized by very high natural variability, superintendents have sometimes been reluctant to accept being able to detect only changes of 50% or more. In some cases, they have understandably been reluctant to be on record as being willing to accept 50% losses without even knowing the loss had happened.

Keep applying common sense tests to all decisions related to trends. In local or regional areas where there are wet years for several years and then several years of dry years, baseline monitoring may have to be done a long time to establish what is normal. Climate and other changes have a way of redefining normal. This fact that should be taken into account when one is trying to detect trends, even if one is using complex methods such as multivariate control charts (see http://www.esajournals.org/esaonline/?request=get-document&issn=1051-0761&volume=014&issue=06&page=1921 and http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc34.htm). How long does the time frame logically need to be to define conditions outside of normal conditions at relatively pristine sites?

For many taxa with large fluctuations in pristine environments, a 10% or even a 20% local change in 10 years would be impossible or costly to detect, and one would not usually go to the trouble and expense if the taxa were not endangered or threatened. Local changes of 50% or even more in 10 years even in pristine sites are perfectly natural for some highly variable species or groups. For extremely variable groups (bacteria, zooplankton, etc.) changes much higher than 50% are normal. Other things being equal,

long term monitoring groups rightly tend to avoid monitoring extremely variable parameters. Although the endangered species criteria above apply globally, they may be of some interest for rough comparison with goals of how big of a change one would like to be able to detect locally, especially when one is dealing with relatively rare or threatened resource.

At the other end of the spectrum, is interesting to compare the kinds of trends that need to be detected for vertebrate endangered species. A new tool is available for trends. An equivalence test has recently been developed for demonstrating the absence of a trend. Sample sizes can sometimes be insufficient relative to the residual variation (and perhaps also autocorrelation) to call a trend. Results from equivalence tests depend critically on the magnitude of the equivalence interval. In one example, a half-life or doubling time of 20 years for population size was discussed for long-lived and relatively stable species. In an example discussed as more appropriate for shorter life-spans and more variable species, a less conservative equivalence region corresponded to a halving or doubling time of 10 years. In an example that used amphibians species on a global (not park) scale, simplifying the definitions of The World Conservation Union slightly, a decline in numbers of >50% in 10 years was said to define an ‘‘endangered’’ species and a decline of 30% in 10 years defined a ‘‘vulnerable’’ species (P. M. Dixon and J.H. K. Pechmann. 2005. A statistical test to show negligible trend. Ecology, 86(7), pp. 1751–1756, http://science.nature.nps.gov/im/monitor/docs/DixonPM_Pechmann_2005_trends.pdf).

Transects are a Special Case:

Transects are often used in long term monitoring and they have appeal for various reasons, but power, sample size, and variance estimates can all be more complex than some alternative simpler designs. For an introduction to statistical aspects, variance estimates, comparison of variance and covariance values, and how spatial correlation complicates the estimates, see Urquhart’s Oregon State ppt discussion slides on stream habitat protocols at http://oregonstate.edu/instruct/st571/urquhart/

.Andrea Atkinson of the SFCN network used similar (to those the section above) but different equations derived from Thompson et al. 1998 (W. L. Thompson, G. C. White, And C. Gowan. 1998. Monitoring Vertebrate Populations. Academic Press, 365 pages) to estimate needed sample sizes to find a 25% change in mean proportions over 5 years in % cover of living coral, using 20 transects per reef (sample size for each transect to estimate the proportion was >250), where variance estimates were based on an analysis of covariance (http://www1.nature.nps.gov/im/units/sfcn/graphics/CoralMonitor.pdf).

There can be more information content in proportions based multiple randomly selected transects than in randomly selected single data points, and going back to the same transects each year increases power (but decreases DF). However, these methods are considerably more complex than most of the others discussed herein, and back transformation introduces bias and should not be done without a good justification. Therefore, before using similar methods, we recommend that all such methods be discussed with a professional, applied statistician to see if they are optimal (vs. simulations and other relatively complex options) after considering assumptions, local factors, and other specifics. The methods that Atkinson used are starting to get beyond

the other comparatively simple ones discussed herein that many quantitative ecologists should be able to perform and fully understand without some outside help.

Confidence Intervals

When estimating confidence intervals (CIs, such as CIs about either side of a mean or median), be a bit suspect of those based on small sample sizes, If sample is under 30, talk to your statistician, or at least calculate sample sizes multiple ways (as listed below) and then adopt the highest answer from the options. Also be a bit suspect of values based on sampling that was not fully representative of the full range of values in the target population in time and space, even when the t-distribution is used instead of the z-distribution.

One common answer to how to calculate sample sizes needed for confidence interval is that the sample size needs to be big enough to get the confidence interval down to a reasonable size (or to a size consistent with project needs). So for proportions a sample size of at least 25 is needed to estimate a proportion of 50% with a 95% confidence interval no bigger than $\pm$ 20% (of a proportion of 50%, see proportion discussions below).

However, simply estimating sample sizes large enough to get a confidence interval down to an acceptable size to define the summary statistic as "well estimated" is not always is not always fully sufficient solution, especially related to confidence intervals about a mean. Null hypothesis tests are used as stand-alones less and less in field studies due to well publicized and very real shortcomings (see discussion on choosing alpha, above).

In this climate there is a tendency to simply calculate a series of confidence intervals and then (in a pseudo-hypothesis test mode) see if the mean and/or data points of one interval overlaps zero and/or the interval of another.

This last one is perhaps the most frequent misconception. CIs may overlap yet there could be a statistically significant difference between the means, see discussion paper by Bower on Minitab Homepage (Some Misconceptions about Confidence Intervals, see http://www.minitab.com/resources/articles/SomeMisconceptionsAboutConfidenceIntervals.pdf).

In other words, too often confidence intervals being calculated for implied uses (such as the pseudo-hypothesis tests) by those with minimal knowledge, even though doing so is often not justified in context and most statisticians would say these types of pseudo-hypothesis tests are not a valid replacement for a proper hypothesis test where both alpha and beta are limited to reasonably small values. In this climate, be especially wary of the most simplistic (one) sample size calculators (those that don't take into account beta), especially for our typically skewed environmental variables. How will the summary statistics and confidence intervals be used? If the summary statistics (means, proportions, etc.) will be used to calculate differences or trends, then steps A-H (above) apply. Also consider the following:

Parametric Confidence Intervals

It takes more samples to estimate a mean adequately (for intended purposes) if variability is high and/or distributions are not symmetrical. Ideally one would also have normal distribution, but for real world distributions symmetry is perhaps more important. Less frequently understood is the fact that properly calculating the sample size necessary for an optimal parametric confidence interval (CI, such as a *t*-distribution confidence interval) on a mean is complicated and subject to more pitfalls than many seem to realize, especially when very small samples sizes (<10 or 25 depending the data) are involved or when folks are trying to stretch the meaning of a CI (for example with the pseudo-hypothesis test discussed above).

One should probably evaluate normality with probability plots for all data between 10<n<25 to judge the advisability of using a normal theory interval. The smaller the sample size, the less symmetrical the data, the less confidence one has the data is from a normal population (and it is difficult to decide from small data sets). With very small datasets, one is usually also less sure that the values sampled represent the full range of conditions of the target population. The more questionable these factors are, the less sure one should be of the validity of calculated parametric confidence intervals. See Part B for additional rules of thumb and detailed discussions.

Sample size calculators typically need a good estimate of the SD, and unless variability in time and space is very low, one does not typically have that in very small samples, especially in the skewed data typical of field and lab environmental variables (and the frequent lack of coverage of the full range of conditions in the sampled population compared to the target population). Again, since many (wrongly) use confidence intervals a bit like (pseudo) hypothesis testing, sample size calculators that test for **"a difference between a mean and a single value"** (AND require both alpha and beta inputs) are probably safer ways to calculate sample sizes for CIs about a mean than the most simplistic of the single-sample size calculators for confidence intervals. Better yet, just use confidence intervals the right way.

Any comparison of 1 or 2 observations to a pre-existing group (a pseudo-hypothesis test use) by seeing whether the new observations fall outside an interval built from that group should be done with a prediction interval rather than a confidence interval. A z-interval should never be used in environmental work, as it assumes that sigma (true but unknown population standard deviation) is known. We never know that. Regarding t-distribution single sample-size calculators, without accounting for power, the t-distribution CI sample-size formula is set at 50% power. So it should be expected that the true interval will be wider than the calculated one 50% of the time. Beta needs to be considered for real-world problems. A good reference on sample size formulas and a sample size estimator for tolerance interval is Kupper and Haffner, 1989. How appropriate are popular sample size formulas? The American Statistician 43, p. 101-105 (Dennis Helsel, USGS, Personal Communication, 2006).

Seemingly recognizing that 50% is not good enough, when discussing the inadequacy of the relatively simplistic t and z equations to estimate sample sizes for confidence intervals, Blackwood stated that the simple t and z statistic formulas that specify only alpha and not beta do not give a reasonable degree of confidence that pre-specified confidence interval lengths will actually be as small as specified [Blackwood, L.G. 1991. Assurance levels of standard sample size formulas (ES&T 25(8):1366-1367], for more details see Part B.

Thus, to assure that confidence intervals are based on adequate size to ensure applicability for expanded or potentially implied uses, avoid calculating sample sizes with relatively simple t-value "single sample" confidence interval sample size calculators with no input for beta. An example of such a calculator would be Zar's equation 7.7 (page 105, Zar 1999, op. cit.). The equations is variance * two-tailed *t*-value, all divided by the d squared, where d is the half-width of the desired confidence interval. Zar admits that accuracy of the equation is not very good at small sample sizes, partly because sample variance is not a good estimate of the true but unknown population variance, and that the equation must be solved iteratively with smaller and smaller sample sizes. This is the same equation used to "estimate a single mean" (in the EPA Version 0.7.2.2) sample size calculator at http://www.epa.gov/earth1r6/6wq/ecopro/watershd/monitrng/tools/sampling.htm. Many (including the Helsel and Hirsch 2002 textbook at http://water.usgs.gov/pubs/twri/twri4a3/pdf/twri4a3.pdf) have explained why these types of simplistic calculators with no input for beta should be avoided and why even those that have input for power should be considered rough estimates for various reasons.

Some might object to calculating power with two-sample calculators when a two sample hypothesis test is envisioned. However, doing so solves some of the problems listed above.

Looking at power does not imply we have to do a hypothesis test, and looking at power is one way to evaluate different monitoring designs including different revisit schedules in panel designs (see Urquhart San Diego Presentation at http://science.nature.nps.gov/im/monitor/meetings/SanDiego_06/SUrquhart_designing_surveys.ppt#423,36,WHY LOOK AT POWER?).

Nonparametric Confidence Intervals

Nonparametric confidence interval (CI) estimates for the median are traditionally computed using the binomial distribution (see description in Helsel, D.R. and R.M. Hirsch. 2002. Statistical Methods in Water Resources. US Geological Survey Techniques of Water Resources Investigations, Book 4, Chapter A3. This textbook is available for free on the Internet at http://water.usgs.gov/pubs/twri/twri4a3/pdf/twri4a3.pdf).

Talk to your statistician before calculating a nonparametric (NP) CI using very low sample sizes. As in the case for parametric CIs, one issue is whether or not the full range of conditions of the target population was included in a relatively low number of samples. In other words, very low sample sizes often increase the probability that the CI will not be especially representative of the target population.

CIs should usually not be calculated if the sample size is less than 6-8. Also with such small sample sizes, the confidence interval is likely to be unacceptably wide for many project goals and/or overlap impossible values (like zero, see further discussion below).

A handy rough first estimate of nonparametric 95 or 99% confidence intervals about a median for sample sizes of 6-120 can be approximated with the UNB tables provided on the Internet at http://erdos.math.unb.ca/~knight/utility/. If one has fewer than 6 observations, trying to do advanced inferential statistics (including CIs) on those few numbers is often unjustified and akin to "much ado about nothing." In other words, one

cannot create substantial new information content where very little information content exists. Trying to read too much into a CI from extremely small sample size (or pretending that a standard error from sample size of 2 or 3 means something), can be a sign that the author either understands little about statistics or is trying too hard to stretch the meaning of anecdotal results.

Once one has moved into the final stages of planning monitoring, if enough representative and credible preliminary data is eventually available, a statistician can help better estimate required sample sizes through bootstrapping simulations. However, keep in mind that "There is considerable controversy concerning the use of bootstrap confidence intervals…Jackknifing and bootstrapping are no remedy for an inadequate sample size. For nonparametric resampling methods, the sample distribution must be reasonably close in some sense to the population distribution to obtain accurate inferences." (1995 W. Sarles SAS discussion on bootstrap confidence intervals at http://www.pitt.edu/~wpilib/bootfaq.html).

Some say that bootstrapping techniques are not recommended unless sample size is over 40 (Elzinga, C.L, D.W.Salzer, and J.W. Willoughby. 1998. Measuring and monitoring plant populations, BLM Technical Reference 1730-1, BLM/RS/ST-98/005+1730, BLM National Business Center, Denver).

Others would say that the caution regarding over 40 is overly cautious and not widely held. The key issue is that a bootstrapped CI based on a sample size of 15 is still a result based only on a sample size of 15. No matter what method for computing the interval is done, it will be worse than one based on n=20. If those 15 don't cover the full range of conditions, the CI will be too narrow. However, bootstrapping is generally recognized as a better way to compute a CI for small samples than standard parametric formulae. So it's the best of a bad situation, but is not necessarily to be totally avoided (Dennis Helsel, USGS, Personal Communication, 2006).

Sample Sizes and Statistics for Taxonomic Richness

This is a complicated subject, see your statistician. For a brief introduction, see Urquhart summaries at http://oregonstate.edu/instruct/st571/urquhart/index.html.

13) Throw Out Other Measures or Strata with Excess Variability

This concept was introduced earlier and is repeated here for emphasis. Once calculations have been done, this topic should be revisited. Measures or strata with excess variability will often prevent detecting a trend or a difference of even a large magnitude with existing budgets. Upon discovering that initial plans for the monitoring design (including what/where/how often to monitor) will not result in being able to detect a difference of concern, adjustments usually must be made. Often monitoring planers throw out measures and strata that are obviously too variable to ever detect an effect size of concern with available budgets. If there is a strong desire to keep the vital sign or measure, try stratifying to get variability and sample sizes down to reasonable levels. If measurement uncertainty is excessive due to poor measurement precision or excess measurement bias, adjust the field or lab methods to get the uncertainty down to

acceptable levels, or choose a surrogate/alternative measure that can be quantified with less measurement-level uncertainty.

14) Optimize Monitoring Plan Details for Affordability and Logic

Monitoring design optimization steps not only include throwing out measures with excess variability, but also considering other steps that could be done to optimize monitoring. For example, if detection probabilities are still too low after the steps above are completed, consider the following:

Often the choice is between: A). monitoring many sites and measures very infrequently and poorly, or B). monitoring fewer sites and measures more rigorously and/or more often. Choice A has too often resulted in data that can be used for very little (if anything) related to management decisions or regulatory purposes. Choice B is sometimes a better option that produces at least some useful information. During plan optimization steps, reconsider the overall affordability and logic of sample sizes, sample placement, sample replication (how many samples at each site, where to sample, how often to sample, when to composite (or not), statistical significance, and statistical power.

The goal is to come up with a combination that will produce acceptable detection probabilities and will produce information (not just data) useful to park managers for resource management decisions. The more (and the earlier) the network quantitative ecologists and statistical consultants can help with these steps, the better.

15) Draft Initial Sample Sizes and Optimized Monitoring Design

Also assemble the best available estimates for input variables (standard deviations, alpha, beta, see list above) to take to the applied statistician (next step).

16) Finalize Sample Sizes and Design with an Applied Environmental Statistician

Once network quantitative ecologists and small groups of specialists that are finalizing protocols and SOPs have completed the steps above, they should strongly consider consulting with an applied statistician, taking that expert the correct information and input variables (above), refined questions (detailed in time and space), and refined target populations. The generic basic design developed for the earlier Phase II report may have been envisioning larger sample sizes and the assumptions may have changed. If the sample sizes have been cut and other changes have been made in design optimization steps taken when developing QA/QC and data analysis SOPs, the revised plan needs to be checked again by a statistician. Typically the first version of chapter 4 (Monitoring Design) of the central monitoring plan is drafted a year or more before the SOPs are finalized. In the next year there have often been disconnects and assumption changes between earlier statistical advice and later changes at the protocol and SOP detail development stage.

Distributions are typically not normal, samples are often not large, and various assumptions may not be defendable. Standard power and sample size analyses may get one in the ballpark but are not optimal in all cases. Some analyses are too complicated to rely solely on plug-in power calculations. Sometimes, multiple hypotheses need to be

considered simultaneously. This requires more complex methods, such as Monte Carlo simulation-based approach to determining sample size and power (see Lukacs Austin Meeting discussion at **http://science.nature.nps.gov/im/monitor/meetings/Austin_05/PLukacs_SampleSize.doc).**

> Remember however, that you need to take meaningful data to the statistician. The initial data available before the start of simulations must have sample sizes large enough (look at the data closer if the sample size is less than 30-200) to be optimally useful in simulations. The initial data must also be relevant and representative of the full range of time and space conditions of the target population. That last caution also applies not only to simulations and other complex calculations but also to simple-algebra Zar sample size calculators.

After completing consultations and final checks with an applied environmental statistician, finalize the following in the protocol narrative and SOPs: 1) sample sizes, 2) minimum detectable differences (or alternative target effect sizes), and 3) sample placement in time and space detail. After monitoring designs are modified and finalized in optimization steps (see chapter 4 discussions in Part B for details), chapter 4 (monitoring design) of the central monitoring plan will also need to be modified to reflect the final design.

The sample size needed to determine a desired minimum detectable difference (and how it was determined) relates to many other issues. Therefore, we suggest networks not only document how MDDs, ESs, and target thresholds were determined in the data analysis SOP, but also include brief recaps or "point to" links in other related sections, such as the discussions of representativeness and completeness in the QA/QC SOP, and the sampling design discussions (Chapter 4 in the central monitoring plan).

17) Estimate the % of Samples That Will Fail

It would be rare for 100% of planned outdoor environmental samples to produce useful data. Seldom are all planned samples successfully obtained and also pass all QC data acceptance criteria. Samples or samplers may be lost in the field, lab or field complications may interfere, and samples can get lost of be spoiled while being shipped to the lab. Weather events may interfere with sampling or analyses (when shipping samples to coastal areas, hurricanes have delayed analyses), staff or equipment failures can be a problem, or delays may cause maximum holding times to be exceeded. A new technician might also use the wrong type of container or otherwise contaminate samples. Therefore, before required sample sizes are finalized, one first needs an estimate of the % of planned samples that may fail. If no other good rationale can be developed, planners sometimes pick a number like 10 or 15% to start with and adjust it as experience is gained.

18) Increase the Planned Sample Sizes Accordingly

Next, adjust required sample sizes upward to correct for the % expected to fail. For example, if 15% of the samples are expected to fail, multiply the required sample sizes developed in 16 by 115%, and edit the plan, protocols, and SOPs accordingly.

19) Include Completeness Goals in a Table in the QA/QC SOP

For each parameter to be measured, include a completeness goal in the SOP. If 15% of the samples are expected to fail, put 85% in the table as a completeness goal.

End of completeness, sample size, and statistics vs. desired conditions outline and chapter.

Missing Values, Useful Data, and Effective Data

The Data Analysis SOP should detail how imperfect data can be and still be used in data analysis or to meet completeness goals. Unless otherwise justified, data that have not met QC measurement quality objectives for precision, bias, and sensitivity are not considered useful and are not included in quantitative statistical analyses. The same is true for: 1) data below qualitative detection limits (MDLs), 2) data between MDLs and MLs (see detection limit discussion further below herein for exceptions, 3) data associated with holding times that have been exceeded or where preservation requirements have otherwise not been met, 3) chemical concentrations where improper containers were used, 4) data beyond minimum and maximum plausible values (checked via range sensibility checks). If some of these will be considered OK for qualitative data analysis, provide the rationale in the data analysis SOP.

How will missing values be handled? This topic is highly related to completeness goals, but decisions hinge not only on what % (like 15%) can be missing, but also on whether or not the missing data is from a critical class of data. For example, suppose the question is "What is the annual temperature?" If all 15% of the data missing are in the coldest part of the year, it would tend to bias the answer.

When using the seasonal Kendall test for trends, an allowance for missing data can be made. In fact, non-parametric tests are sometimes chosen because 1) they are not affected when the distribution of data is not normal, 2) are insensitive to outliers, and 3) are less impacted (or not impacted) by missing or censored data (Harcum, J.B., J.C. Loftis, and R.C. Ward. 1992. Selecting trend tests for water quality series with serial correlation and missing values. Water Resources Bulletin 28(3):469-478). The decision of what percent of the data can be missing and still pass completeness goals should include a common sense check relative the questions to be answered and the statistics to be used.

Include this topic in discussions with your applied statistician, since imputation options tend to be complex. See Part B for additional discussion.

Part B also contains generic definitions for useful data and effective data. Depending on the questions and project data quality objectives, data from screening methods can be "effective" even it was not collected by the most precise or accurate methodology (see detailed discussions of effective data in Part B.

Although the data analysis SOP should cover the basics of what will be done with missing data, more detail, along with how values below detection limits (see discussion further below related to low level detection limits) will be handled, should be covered in the QA/QC SOP.

QUALITY CONTROL:

Data Comparability (Internal/NPS and External/Other Regional Data)

We are now moving from QA topics to QC topics. Comparability is usually considered a QC basic, albeit one assured qualitatively. More statistical tools are now being developed, and in future years the comparability may be assured more quantitatively.

For Internal Data Comparability: What will be done to maximize temporal and methodological consistency in NPS data? Control typically involves limiting changes in internal NPS methods or timing of sampling to help insure our own newer data is comparable to our older data. However, due to advancing technology and other factors, changes in both methods and personnel are inevitable. When such changes occur, any resultant measurement bias from the change should be documented in the Cumulative Measurement Bias SOP. The question then becomes: Is our internal NPS data from both old and newer measuring systems comparable enough that the different sets of data could be combined for purposes of determining trends or making management or regulatory decisions? If not comparable enough for that purpose, newer data and older data will often need to be normalized to data as of one (baseline, often starting) date. For more information, see section below entitled "Include a Cumulative Measurement Bias SOP,"

For External Data Comparability: What will be done to achieve comparability with other regional data sets, such as those of USGS, states, NOAA, or CERCLA sites)? Again, what will be done to insure our NPS data are comparable enough to the data from other state and federal agencies that need to be convinced our data is credible and comparable, given our purposes for monitoring? Is our NPS data comparable enough to other important outside data sets that the two sets of data could be combined for purposes of determining trends or making management or regulatory decisions? Has the chemical lab proposed for use 1) passed federal round robin blind sample checks (see FWS example at http://www.fws.gov/chemistry/acf_how_we_select.htm), or 2) performed acceptably (to NPS WRD) in other federal round robin blind checks (see USGS example at http://bqs.usgs.gov/srs/), or 3) been approved to work for the parameter of interest and media of interest by the Federal National Environmental Laboratory Accreditation Conference (NELAC, see http://www.epa.gov/nelac/)? Labs that participate in the USGS SRS round-robin comparison are found at http://bqs.usgs.gov/srs/EnrolledLabs.xls. Checking the BQS site allows networks to assess QC performance for the lab in question lab for the analytes in to be analyzed (BQS emphasizes water column parameters

including nutrients). Networks can then decide if the performance was acceptable for project data quality objectives. The most recent performance from labs is at http://sedserv.cr.usgs.gov/srs_study/reports/index.php. For those enrolled labs that USGS has not numbered (xls file above), the information is private and can only be obtained from the lab.

A final check should be made to make sure both the lab and the field method SOPs attached to the protocol are detailed enough to allow for reproducibility of exactly the same methods by third parties. Are they also detailed enough to allow judgments about the comparability of the data with the data of other agencies? Perfectly comparable data can be merged and analyzed together without introducing problems.

These issues are just as important for biological monitoring as for water chemistry monitoring. Interagency efforts are now being made to come up with acceptance criteria to determine data comparability.

Comparability in Agreement or Pass/Fail Scores

As of 2006, there are no universally accepted ways to assess "agreement" (a different topic than correlation) in ratings or scores for biotic "condition." Highly correlated scores (such as index of biotic integrity scores) indicate high "association" but do not guarantee a strong strength of "agreement." For example, results from one state sampling protocol may rate stream condition consistently one level higher or lower than that of another state or federal program. In this case, the strength of agreement is not strong, although the correlation/ association may be very strong. Accordingly there has been much recent interest in various options for measuring strength of agreement.

It is easier to make comparisons with scores or ratings when there are only two choices rather than with many different categories. These matters are the subject of much current interest. Again, it is much easier to make comparisons with two variables or two scores than with many. Networks should probably consider looking at such issues from different angles, including some relatively simple and intuitive ones.

It helps if dichotomous decisions can be made. Even in a relatively complex rating system (very poor, poor, fair, and good) the key goal might be maintaining "good" condition. Simple calculations could be made relative to the % agreement where both methods resulted in "good" scores, the relative percent difference between two scores, or the % bias comparing results of one method to another.

As another example, what % meets quality standards (pass/fail)? With enough data, networks can usually quantify the proportion of river miles that either passes or fails water quality standards. These values can sometimes be compared with other category dividers (say between fair and poor) in more elaborate systems.

. This approach would be consistent with a reality-check step of looking at the issues from different angles, including some relatively simple and intuitive ones. In other words, use intuitive and simple lines of evidence in addition to exotic coefficients (such as kappa) when possible

When there are multiple ratings (very poor, poor, fair, and good), things get more complex. Some have suggested using kappa or weighted kappa to look at agreement of IBI scores or to evaluate agreement in ratings of stream condition. For example, the EPA summary of the Mid-Atlantic Integrated Assessment Maryland case study took this

approach. They also looked at agreement via kappa and several other ways (multiple lines of evidence, see http://www.epa.gov/maia/html/biomd2a.html and http://www.epa.gov/maia/html/biomd2b.html).

The free Internet McBride Cohen's kappa calculator can be used as one way to look at agreement of dichotomous data (http://www.niwa.co.nz/services/statistical/kappa). McBride also has a calculator on the net to calculate Lin's "concordance correlation coefficient." This statistic appears "to avoid *all* of the shortcomings" associated with the usual procedures (such as a Pearson correlation coefficient) and can be used as one way to look at strength of agreement of continuous data (http://www.niwa.co.nz/services/statistical/concordance).

However, kappa and similar methods tend to be complex and they have their detractors. Detractors say kappa is over-used. They also point out assumption complications and that not everyone agrees on how high a kappa score has to be to reflect various gradations of agreement. So agreement coefficients are something to have your applied statistician approve before finalizing them for particular applications. For more detail, see Part B.

Why Document Quantitative Quality Control?

We are now moving from more qualitative subjects to documenting quantitative measurement quality objectives for quality control (QC) data quality indicators. Before doing so, we will first answer the following question:

> Why do we need to quantitatively control and document QC data quality indicators like measurement sensitivity, measurement precision, measurement systematic error/bias, and blank control?

The short answer is that most states and regulatory agencies require us to do so as part of a required quality assurance project plan (QAPP) and/or due to state credible data laws. It is also required by NPS WRD and the VS checklist.

If QC measurement quality objectives are not controlled and documented, most regulatory agencies will not consider our data to be credible enough to use in Clean Water Act decisions or processes (like section 303d use impairment determinations and Total Maximum Daily Loads/TMDLs). For a summary of the Clean Water Act, TMDLs, and 303, see BLM website at http://www.blm.gov/nstc/WaterLaws/abstract2.html).

It is expensive to collect data, so to the extent possible, data collected should be useful for multiple purposes, especially purposes that will help resource protection or management, or be helpful related to GPRA goals.

Even if no outside entity is "making us" do so, there are logical reasons why we should control and document the performance of QC data quality indicators:

1. The scientific community has known since the 1930's that measurement processes need to be quantitatively controlled to produce credible data, so not controlling the measurement process would be difficult to justify in today's world.

2. Limiting problems with measurement precision and limiting measurement bias to specified-amounts; together control how imperfect the measurement process can be. Without such controls, we have no way to estimate how badly the measurement process is performing. Measurement uncertainty could be extremely large, and we wouldn't know it.
3. Accordingly, if ever challenged in court, not controlling and documenting measurement performance for QC indicators (measurement sensitivity, precision, bias, etc.) would be impossible to defend. Upon challenge, we would not be able to document boundary limits on the magnitude of measurement uncertainty.
4. This is long term monitoring, and controlling and documenting data quality indicators gives our data a better chance to be considered credible in future years. In future years, we would not want our data thrown out because someone had then decided that all data without full quality control documentation was unacceptable and would not be used. Judging data without QC documentation as not credible or useful is already happening and the tendency to do so can be expected to become more common.
5. In long term monitoring, method, SOP, and staff changes are inevitable. Documenting changes in measurement performance through such changes is therefore even more important than for short term projects. Documenting QC performance of the old and new methods helps one determine whether or not a change in values was the result of a true change in the environment or whether it was the result of something else.
6. NPS WRD requires that all VS or WRD funded water quality data are archived in Modernized STORET. STORET contains fields for detection limits, precision, systematic error/bias as % recovery, blank control results, etc.
7. NPSTORET is even more insistent on having detection limits. In NPSTORET, if one enters a result with detection conditions "Not Detected", "Present, below Quantification Limit", or "Present, above Quantification Limit", one will not be able to enter a value. Data with these conditions requires entry of the relevant limit.
8. We need QC results to be able to interpret data in a common sense manner. For example, we need low-level measurement sensitivity results to understand whether or not the analyte is present and how big of a change in the concentration of the analyte is believable as a real change, rather than a random error in the measurement process.
9. We logically agree with the DOI information quality guidelines that emphasize QA/QC basics. These include ensuring and maximizing the "quality, utility, objectivity, and integrity of the information." Another basic is to ensure "a high degree of transparency" about data and methods used to generate the data. Optimal transparency would include admitting that is data are not perfect but that the degree of imperfection could not have exceeded MQOs specifications for precision, bias, and sensitivity.

10. Reproducibility is not only a QC basic, it is a "sound science" basic. Unless one documents measurement performance characteristics, it will be very difficult for another party to reproduce the result independently.
11. Programs that have instituted improved quality control over the years see their QC scores improving, even for field measures (see USGS example at http://fl.water.usgs.gov/Abstracts/ofr98_392_stanley.html). That document is for the period ending in 1997. The current USGS field measurement QA check program is at http://nfqa.cr.usgs.gov/. The advantage to agencies using these kinds of checks is that if scores go downhill, the agency would know to make corrections until the situation was better. If there were no checks, an agency would not even know the measurement process had deteriorated.
12. Good quality control is necessary to scientific credibility, and without scientific credibility, a long term program would be rightly more susceptible to budget cuts.
13. Good QC plus good documentation of measurement changes that bias scores up or down (see separate discussion herein) makes finding long term trends in a defendable way possible.

For those who might still believe that journal-style post-project peer review is all that is needed, take a look in Part B at discussions of why that solution (although one helpful step) is not complete by itself. Among the reasons: 1) it is too late to change the design or outcome, 2) scientific journals that have studied there own peer review processes have found glaring inadequacies, 3) many projects that have been published in peer reviewed journal articles have no documented QA/QC at all. A 2006 summary of some of these issues including why past (post-project) peer review does not work well to insure information or data quality; and related resources is at http://www.info-cybernetics.org/KCPR2006/website/default.asp**.**

Measurement Sensitivity

When measuring water quality chemicals at very low levels, a system that can accurately measure and detect a very low concentration is more sensitive than one that can only detect the presence of the analyte at higher concentrations. The more sensitive the measuring system, the lower the low-level detection limits are.

Toxic chemicals can be hazardous at very low levels, and there is typically a concern about whether or not they are present in parks, even at very low levels. Likewise, some pristine waters in the NPS have very low concentrations of nutrients, and the parks want to keep them that way. Both of these scenarios lend themselves to documenting and controlling measurement sensitivity with the lowest-practicable detection limits. Low level detection limits have been the most common way sensitivity has been handled in the past, and for water quality parameters sometimes present in very low amounts, they are still critical.

Even if we are always measuring in higher measurement ranges (well above the low-level quantitation detection limits) we still need to control measurement sensitivity. For some parameters measured in the field (pH, temperature, conductivity, biological

observations, physical habitat observations, etc.), one seldom, if ever, encounters extremely low levels. In these higher measurement ranges, the smaller the (true) change that a measuring system can accurately detect, the more sensitive the measuring system is. For these types of measurements or observations, low level detection limits are less relevant and/or less helpful, and an alternative measurement sensitivity (AMS) method to estimate measurement sensitivity is explained farther below.

Low Level Detection Limits (MDLs and MLs)

In 2004, EPA clarified that "detection" indicates the presence of a pollutant in a sample. Quantitation, on the other hand, indicates how much pollutant is in the sample (http://www.epa.gov/waterscience/methods/det/#proposed).

Minimum Requirement: In the QA/QC SOP, list (a table is fine) pre-project targets (and how often they will be estimated once monitoring begins) for the following low-level detection limits:

1. A **semi-quantitative** method detection limit (MDL) and
2. A **quantitative** (minimum level of quantitation) detection limit (ML).

For NPS standardization, the MDL and ML are the suggested defaults. We suggest a minimum frequency for calculation of these values of at least once a year or when there is a change in the measurement process. Using the default suggestions given in many EPA methods would be acceptable. Many EPA methods suggest that MDLs should be determined every six months, when there is a change in a measurement instrument's response or in background response, when a new matrix is encountered, or when the lab believes (for whatever reason) that there may have been a change of low-level measurement sensitivity. For example, when a new operator begins work or when there is a significant change in the measurement process (new method or new instrument), one should suspect that sensitivity may have changed and therefore recalculate MDLs. Some labs take the precaution to calculate MDLs more often, and that fine. It exceeds our minimum suggestion.

As first mentioned in the section on picking methods and SOPs, there should be an emphasis on picking methods and labs able to achieve a semi-quantitative MDL detection limit lower than the lowest water quality standard (including chronic standards) or other threshold benchmark of concern (optimally at least 1.6 to 2 times lower).

The MDL and ML are to be calculated as follows:

MDL:

List the target standard EPA method detection limit (MDL), for each parameter to be measured, in a QA/QC SOP. Labs should be instructed to calculate the MDL as explained in Appendix B to 40 CFR Part 136—Definition and Procedure for the Determination of the Method Detection Limit—Revision 1.11 (http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&sid=fadf72a44c330e5dd957e61dddab97e7&rgn=div9&view=text&node=40:21.0.1.1.1.0.1.6.2&idno=40). That same definition was reiterated, further explained, and

defended by EPA in 2004 in their very lengthy (200+ pages) updated recommendations and discussions of pros and cons of alternative detection limits past and present at http://www.epa.gov/waterscience/methods/det/rad/rad.pdf.

> To determine the MDL, at least seven replicate samples with a concentration of the pollutant of interest near the estimated detection capabilities of the method are analyzed. The standard deviation among the replicate measurements is determined and multiplied by the upper (one sided) t-value for n-1 degrees of freedom (in the case of 7 replicates, the multiplier is 3.143, which is the value for 6 degrees of freedom).

Although most labs and even some EPA staff and published EPA methods do not always use all the steps suggested by EPA to calculate an MDL, most at least eventually use the central equation of Method Detection Limit (MDL) = t times S, where, t = the one-sided Student's t-value for seven replicate (precision repeatability) samples. In this equation, for 7 replicate samples, t = 3.143, so MDL = 3.143 times the sample (n-1 version) standard deviation for the 7 replicate measurements of a blank.

The same equation has used in Standard Methods Book (http://standardmethods.org/) definitions of an MDL, and by many states and others. The MDL is usually said to be the lowest-value we really believe with 99% confidence is different than zero. Different than zero calls for a one-sided statistical comparison, which is why we use the one-sided *t*-value. The same EPA equation is usually used to estimate estimated detection limits (EDLs), but with fewer steps than one uses in estimating an MDL. Thus MDLs and EDLs are not usually the same value. Calculating EDLs is most often a preliminary step on the way to estimating MDLs. EDLs are often calculated with low-level standards or solutions rather than blanks.

When blanks never produce detectable signals, it is common to estimate MDLs in samples with the lowest possible levels where a signal can be detected (EPA, 2004, http://www.epa.gov/waterscience/methods/det/rad/rad.pdf).

To avoid confusion, alternative semi-quantitative detection limits (EDL, LOD, IDL, LLD, etc., see Part B for details) should ordinarily not be listed. The exceptions include:

1) If a USGS lab is used, the USGS alternative to the MDL (the LT-MDL) can be listed along with how and how often it is calculated (usually using a f-pseudosigma rather than a standard deviation and sample size higher than 7 (see **http://water.usgs.gov/owq/OFR_99-193/)**, or
2) If there is no good way to calculate or find an MDL but an estimated detection limit (EDL) can be found or logically calculated (sometimes the case for bacteria or chlorophyll) and calculated, the EDL can be defined and used. Some labs basically let electronic instruments define detection limits, since some instruments censor low level values in the noise range (below calibration curve limits). If this is the case, exactly how the semi-quantitative detection limit was determined, and why standard MDL calculations were not made, should both be documented in the sensitivity section of the QA/QC SOP.

NEMI (www.nemi.gov) sometimes gives the lower end of the calibrated range as a "range-derived" lower detection limit, and this or some other rough estimate of a MDL might be used in QC tables when bad precision when measuring close to an MDL does not require a more stringent calculation of a proper low-level MDL.

In cases where one is always two or more times above the MDL and never encounters really poor precision from measuring too close to a MDL, alternative measurement sensitivity (AMS) should be periodically calculated and reported (see AMS discussion farther below).

Minimum Level of Quantitation (ML)

Unless otherwise justified, define and calculate the minimum level of quantitation detection limit as 3.18 times the MDL. The resulting value is the same as the (low level) lower quantification limit (LQL, a STORET-specific term), but it is suggested that the ML terminology be used rather than LQL except when dealing with STORET. In previous versions of Part B, this detection limit was referred to as a PQL, but 2004 EPA guidance documents make it clear that the ML (as 3.18 times the MDL) is a better default term for the quantitative limit. The ML is typically a bit lower that the Currie-defined limit of quantification (LOQ). In 2004, EPA provided extensive documentation of where 3.18 came from and how it relates to generic detection concepts, see details in EPA 2004 at http://www.epa.gov/waterscience/methods/det/rad/rad.pdf). A more elaborate definition of the ML including rounding rules is also contained therein. The 2004 EPA document also explains that the ML is "the lowest level at which the entire analytical system must give a recognizable signal and acceptable calibration point for the analyte. It is equivalent to the concentration of the lowest calibration standard, assuming that all method-specified sample weights, volumes, and cleanup procedures have been employed.

As explained in EPA 2004 (op. cit., above), some of the criticisms of the MDL and ML relate to single lab vs. multi-lab comparisons. If a network needs to do so because it is dealing with multiple labs or cannot achieve quantitative detection limits at 3.18 times the MDL, the network could alternatively define a multi-lab PQL as 5 (rather than 3.18) as suggested in the standard methods book (op.cit.). Just make it clear that the PQL detection level is a multi-lab achievable rather than a single lab quantitative limit.

In the context of a single lab, there are disadvantages for using 5, and it should be a justified the exception rather than a default choice. Using 5 would result in being able to report fewer low level values (see following section). Also, some EPA methods specify the use of 3.18. Top experts in the field now consider 3.18 to be sufficiently high to protect against false negatives, to be the value most commonly used, and to have other advantages over 5 (for details see Part B, or D. Helsel. 2005. Nondetects and Data Analysis: Statistics for Censored Environmental Data. Wiley. 288 pp., http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471671738.html).

To avoid confusion and for NPS standardization, use the ML rather than alternative quantitative detection limits phrases or acronyms. It is easy to become confused in the alphabet soup of the great many alternatives. Alternatives include practical quantitation limits (PQLs), minimum reporting levels (MRLs), reporting levels (RLs), limits of quantitation (LOQs), minimum quantitation limits (MQLs), sample quantitation limits (SQLs), Contract-Required Quantitation Limits (CRQLs), or an interlaboratory quantitation estimate (IQE). These and many other variations are

explained in EPA 2004 (op. cit.). Instead of using these terms, convert all such quantitative limits to MLs.

There are two exceptions where terms other than the ML can be used:

1. When dealing with STORET, the phrase "lower quantitation limit" (LQL) can be considered a synonym for a ML.
2. If a USGS lab is used, the USGS alternative to the ML, the laboratory reporting level (LRL) may be used instead of the ML. The LRL is defined as two times the USGS LT–MDL (http://water.usgs.gov/owq/OFR_99-193/level.html). If the network is going to use USGS LRLs, explain how and how often they will be calculated and reported.

No matter what quantitative detection limits are used, how they are calculated should be explained in the sensitivity/detection limit part of the QA/QC SOP. Once monitoring begins and new data is put in STORET, it should also be put in STORET metadata, as explained in the next section.

For toxic chemicals, it is particularly important that the lab can achieve ML quantitative detection limits that are below the benchmark, water quality standard or criteria, or other threshold levels known to be associated with harmful effects.

How Will Values Below the MDL or ML be Reported and Analyzed?

The measurement sensitivity/detection limit section of the QA/QC SOP attached to each protocol should explain how data below any of the listed detection limits will be handled, not only for reporting into data bases, but also for data analyses. Along with how missing values will be handled (see discussion further above related to completeness), how values below various detection limits will be handled should also be covered in the data analysis SOP.

An acceptable option (and one already adopted by some VS networks) is to state that the recommendations in the recent Helsel Book (D. Helsel. 2005. Nondetects and Data Analysis: Statistics for Censored Environmental Data. Wiley. 288 pp., http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471671738.html) will be followed. Some of Helsel's recommendations, such as "Why not substitute one-half the detection limit for nondetects?" are summarized in relatively plain language at http://www.practicalstats.com/. Issues related to storing nondetects in databases are summarized briefly at http://www.practicalstats.com/News/Autumn04.pdf.

NPS data needs to go into STORET, and the Helsel book (op. cit.) considers the modernized STORET default recommendation for writing to a database to be fully acceptable. Therefore, we are adopting this as a default NPS recommendation. Modernized STORET and NPSTORET both suggest that we not report into a database any value higher than the MDL but lower than the ML. Instead, the detection condition field is set to "Present, below Quantification Limit." With that detection condition, STORET automatically enters "*Present <QL" in the result field. A major advantage of this approach is that no "estimates" are treated as quantitative when they are not quantitative. NPSTORET is consistent.

In (eventual) statistical analyses, values between the MDL and ML are best interpreted using either an interval-censored method (parametric) or a rank-based method (nonparametric). In the latter, all in-between values are represented as the same tied rank. The older recommendation of censoring to half the MDL is clearly no longer recommended.

For reasons consistent with both STORET and NPSTORET rationales, Helsel recommends that numerical values not be reported into data bases if the values are below the MDL or the ML, and that one should not report nondetects as half the detection limit. One should also not report nondetects as a negative (minus or -) sign followed by the actual MDL value, because someone invariably decides it really is a negative number (Helsel, op. cit.).

These recommendations are all followed in NPSTORET, which will not allow entry of values below the ML. The MDL and ML limits are entered into NPSTORET, and by using STORET detection condition coding results, one can find out how many values were below the MDL or between the MDL and ML.

Values above the ML are classified in EPA's modernized STORET database with the detection condition of "Detected and Quantified" This is ideal, and according to EPA STORET Staff, this is optimally the only choice which permits reporting a single number.

Although not recommended in the Helsel book (op. cit.), for the special case of NPS analyses of "precautionary principle" comparisons with standards or criteria, one might choose to censor all data below the ML to the exact value of the ML, but that is only a very special (worst-case, trying to be very precautionary and totally avoid false negatives) example of a data analysis strategy, and one would never substitute the value of the ML in a long term network storage data base field for measured concentrations.

Alternative Measurement Sensitivity (AMS)

In a sense, the MDL and ML are commonly used strategies for bounding, and thereby controlling, measurement uncertainty and measurement sensitivity when one is trying to measure extremely low values, at or near the lowest-value that can reliably be measured (the ML), or the lowest-value that can be distinguished from zero (the MDL, see discussions above).

When one is consistently measuring above the ML in the normal quantitative measurement range, one can use standard National Institute of Standards and Technology (NIST) methods and terminology.

NIST works with its international counterpart, the International Organization for Standardization (ISO), to standardize measurement methods and terminology used in science and engineering worldwide. They have reached consensus on measurement basics. These include:

1. No measurement is perfect. Each is an approximation.
2. Individual measurement data points are not complete unless accompanied by a statement about the uncertainty of that approximation.

NIST has acknowledged that standard NIST/ISO methods to calculate NIST "expanded measurement uncertainty" are not applicable for very low (below quantitative-ML detection limits) ranges of measurement (N. Taylor and C. E. Kuyatt. 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297, http://physics.nist.gov/Document/tn1297.pdf).

However, if one is consistently measuring values well above the very difficult low-level ranges, expanded measurement uncertainty (our AMS when sample size is 7 and confidence level is 99%) is very helpful in estimating quantitative-measurement-range measurement sensitivity. In some cases (pH and temperature, for example), no blank or other zero-value solutions are available. In other cases, zero-value solutions would not be relevant to the measurement ranges of interest.

When measuring conductivity in the field or when making biological or physical observations in the field, a low level detection limit is usually not very helpful. However, in all of these cases where standard MDLs and MLs are difficult or not relevant, controlling measurement sensitivity is a still a QC basic that should not be ignored.

Since we are not considering the lower end of the range in which one cannot use NIST/ISO standard calculations for expanded measurement uncertainty, there is no reason not to do things the standard way. The APHA (Standard Methods Book, http://standardmethods.org/) also recommends that expanded measurement uncertainty be calculated the same way as NIST and ISO, lending even more credibility and standardization to the method.

Therefore, the NPS default suggestion is to calculate an Alternative Measurement Sensitivity (AMS) based on NIST expanded precision uncertainty using a sample size of 7 and 99% probability. This satisfies two needs at the same time: 1) the need to control measurement sensitivity when in the normal quantitative measurement range and 2) the need to have a plus or minus value to put in the STORET "analytical procedure description" text box.

The reason we are calling this estimate "alternative measurement sensitivity" is to distinguish it from standard low-level MDLs and ML estimates of sensitivity.

To calculate AMS in a way that is as functionally analogous to sensitivity based on MDLs and MLs as possible, one should use the same factors used in MDL and ML calculations: 1) The same number (seven) of replicate measurements of one sample and 2) the same (99%) level of confidence.

It may seem like the discussion has been shifted from sensitivity to precision. However, measurement sensitivity is usually derived from looking at precision results in an especially rigorous way but doing so less often than one samples normal precision QC samples. That is what we are suggesting here.

To determine AMS sensitivity, measure one typical-concentration sample 7 times and use a 99% level of confidence. For contrast, when one controls precision separately (and more often), one measures a single QC precision sample twice (not 7 times), and one doesn't worry about 99% confidence, as explained in the next chapter.

The terminology AMS is used to emphasize that what we are dealing with here is typical measurement range measurement sensitivity. We are not dealing with 1) standard QC precision samples, 2) MDLs, 3) MLs, or any form of lowest-level measurement sensitivity.

AMS results are not recorded in STORET detection limit fields. Instead, they should be recorded in the STORET metadata "analytical procedure description" text box. AMS should be calculated in the same manner that NIST/ISO/APHA "measurement precision expanded uncertainty" is calculated, using 7 replicate samples and a 99% level of confidence (for details, see N. Taylor and C. E. Kuyatt. 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297, Web at http://physics.nist.gov/Document/tn1297.pdf).

A plain language step-by-step for doing this is provided as follows: first, one samples one NORMAL (not a blank as was the case for MDLs or MLs) environmental sample seven times. Those 7 results are used to calculate the sample standard deviation. Next the sample standard deviation times is multiplied by 3.708 (the 99% confidence middle t-value for sample size 7) to get expanded precision uncertainty = AMS. Note: one can confirm and calculate both middle (two-sided) and left (one-sided) t-values in t-tables in statistics books, with statistical software, or with the SurfStat Australia calculator at http://www.anu.edu.au/nceph/surfstat/surfstat-home/tables/t.php.

The result of this calculation is functionally analogous to a ML but uses the two-sided (middle) t-value because we are not only interested in one side (distinguishing a value from zero) but in the two-sided issue of how large of a difference between two individual values we can justify as actually being a true difference (from one another) rather than being due to measurement process noise (random up and down imperfections in the measurement process).

The AMS (as expanded precision uncertainty) should be calculated and reported no less than at least once every sampling season, and more often until a reasonably consistent range is developed. Like MDLs, AMS should be re-calculated when something significant in the measurement process changes. For example, if the person measuring, the measuring instrument, or the methods or SOPs change, recalculate AMS to see if measurement sensitivity has changed.

The AMS result may be used not only to fill in blanks in STORET but also to bound measurement uncertainty on each single data point. This allows us to be consistent with modern (and transparently honest) scientific thought that no measurement or observation is perfect. Instead each is simply an estimate. Just as confidence intervals express the uncertainty about a mean, there is an AMS interval of uncertainty around each single data point.

If measurement precision and sensitivity were very good, a result for a single data point might be something like 45.676 plus or minus 0.003. In the more common (for field monitoring) scenario where neither sensitivity nor precision are that good, the result for a single point might be reported as something more like 50 plus or minus 30.

Either way, the result can be entered into STORET in the plus or minus field for "precision" in the CHEMICAL DATA RESULT ENTRY BOX. Bounding uncertainty in this way is a more modern alternative to using rounding rules to decide how many significant figures one should carry in final result (after all multiplications and divisions or other manipulations have been completed).

Unlike the MDL or ML, historically data has typically not been censored based on AMS/measurement precision expanded uncertainty values. However, as a statistical analysis strategy, one could take the worst case end of the range. For example, suppose the highest pH value considered safe was 9.0 and the only piece of information available

was a single value measured was 8.9 plus or minus 0.3. Single values are anecdotal, but one might say the single value could be as high as 9.2 and therefore may exceed the criteria.

Many biological inventory and monitoring projects have not historically estimated measurement sensitivity. However, there is usually no reason why one person could not calculate AMS after measuring one sample 7 times (or perhaps have one sample measured by 7 different biotechs, or something similar). It may take some ingenuity in difficult cases. In the case of destructive sampling, it may require a sampling nearby areas rather than re-measuring one identical sample. In the same way we study precision plus (see "precision plus" discussion in precision chapter below), we may need to estimate "AMS plus" in some cases, where measuring one sample repeatedly is impossible. As long as an AMS plus estimate reflects are very small amount of variability, the sensitivity of the measurement process is also well controlled. If AMS plus is high, more study would be needed to find out if the extra variability is from the measurement process or from potential true variability of near but not-identical samples (the plus part).

With careful thought, it should usually be possible to develop a common-sense way to adapt the AMS/expanded measurement precision uncertainty logic into AMS functional analogs for various types of biological monitoring. The key is to try do so in a way that "makes sense" while still addressing the issue of logically estimating and controlling measurement sensitivity and uncertainty.

Can one list both low level detection limits and alternative measurement sensitivity for field measurements? Yes, as explained above, there are separate places in STORET to put both results. Whenever one may encounter very low concentrations, it would be optimal to do both.

Is it OK to use the lower end of the applicable measurement range (in manufacturer's specifications) as an estimate of the MDL and for reporting to STORET? No. This is not ideal and should never be done if the lower end of the specification range is zero. Doing this would never be recommended where true MDLs are needed (for toxics and very low-level nutrients in very pristine lakes). In these cases, proper MDLs and MLs should be calculated (see discussion above). Zero is never a correct answer.

What if one is always measuring values well above the lower end of the range and does not encounter really bad precision? This is the scenario where an AMS is appropriate. In this scenario, one would expect to be able to meet precision MQOs. Bad precision would be indicated when the field instrument will not settle down on one reading but just keeps changing, even after a reasonable period has been allowed for the instrument to settle down. Really poor precision may be an indication of an instrument or calibration problem, but it can also be a clue that one may unknowingly be measuring values no greater than 2 x an estimated MDL. If it turns out that one is sometimes measuring that close to an MDL, estimating a proper MDL would be appropriate.

MDLs and MLs are not required in STORET and are only required in NPSTORET if one reports the detection conditions of either "Not Detected" or "Present, below Quantification Limit." So, if one never encounters this scenario, one need not enter a MDL or ML into NPSTORET (or in STORET). In this case, one should simply calculate and report a more appropriate type of sensitivity (AMS, as explained above).

What does one put in the MDL and ML tables in the QC SOP in this situation? In many cases one can use another program's defaults (for example, a state or EMAP's

MDLs). One might also just enter a code in the table that is explained in more detail at the end of the table. The code explanation might explain that as long as precision MQOs are met above twice the MDL and the field readings settle down to one value, these other-agency MDLs and MLs were considered sufficient. MDLs (or other forms of measurement sensitivity) need not always be the most stringent ones available, but they do need to be listed and meet project goals.

Resolution

Measurement resolution relates to a single measurement value and is usually somehow related to the fineness of the measurement scale, but beyond that, little seems standardized. It is typically not acceptable to use the resolution specifications of the manufacturer of a field meter for AMS, a low level detection limit like an MDL, or for precision. There are hints that what many meter, probe, or sonde manufacturers call resolution is perhaps most like measurement uncertainty/AMS, but the lack of consistency between manufacturers regarding how resolution is estimated, prevents us from calling resolution a synonym for AMS or measurement uncertainty.

This lack of consistency is one reason that the word "resolution" is not typically seen in environmental quality assurance project plans (QAPPs). Resolution specifications often do not correspond well to real-world field precision or sensitivity. Sometimes, resolution specifications seem to have been developed at least partly for competitive advantage in ideal lab situations. So, to document measurement sensitivity in the actual environment being monitored, one still has to measure actual field performance for measurement uncertainty/AMS or low-level detection limits like MDLs.

Therefore, the word resolution should usually not be used in planning water monitoring. Those who have used the word are too often talking about other more commonly understood QC concepts, such as precision, uncertainty in accuracy, sensitivity, detection limits (a special case of sensitivity when signals are low), or measurement uncertainty/AMS. Since these concepts are defined in detail separately herein and more universally understood, there is typically no need to address the concept of resolution separately in water quality or contaminants QAPPs or QA/QC SOPs.

Certain GIS/Remote sensing and non-linear biological categorization applications may be exceptions. In remote sensing, the word resolution is often used for a concept more broadly recognized in other disciplines as sensitivity. In any case, if the word resolution is used, how it is estimated should be defined in detail.

Measurement Precision as Reproducibility and/or Repeatability

Measurement precision (actually imprecision, but according to tradition and common practice most ignore that) is the variability of repeated independent measures of the same object**.** ISO defines precision as "the closeness of agreement between independent test results obtained under prescribed stipulated conditions." Contributors to lack of perfect precision include random (up and down) measurement error and random sampling error. Unlike systematic error/bias, precision does not depend on the true or expected value.

NIST clarifies that the "prescribed stipulated conditions" should include documentation about whether the precision is "precision under repeatability conditions" or "precision under reproducibility conditions" (B. N. Taylor and C. E. Kuyatt. 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297, at http://physics.nist.gov/Document/tn1297.pdf).

In more plain language, precision is simply the variability of the different observations compared to each other. For sample size two (a duplicate), in water quality work, precision is most commonly expressed as a relative percent difference (RPD). For larger sample sizes, precision performance is most often reported as a relative standard deviation (RSD or coefficient of variation expressed as a percent). Precision performance can also be expressed as a sample (n-1 version) standard deviation, though this has routinely not been done in water quality or contaminants work.

Precision is not only estimated in a rigorous way every now and then to estimate sensitivity/AMS/measurement uncertainty (as discussed above), it is also estimated much more often in a much less rigorous way (sample size 2 rather than 7, using typical concentration environmental samples rather than blanks) to monitor measurement precision. This is a standard QC requirement and is the subject of this section. Precision duplicates are typically done often enough to document that the measurement process is remaining "in control." When controlling precision in batches or groups, QC precision samples are usually collected every 20 samples or so and then measured twice (duplicates) to control measurement precision on a regular basis.

A check list of items to be included in a separate QC SOP for each protocol is provided as follows. For each measurement done in the field or lab, are the following adequately covered?

- A measurement quality objective (MQO) for precision, usually $\pm$ percent.
- Will the MQO be used as a data acceptance performance standard? It usually should be an acceptance criterion.
- What is the data comparability source of the MQO (state, USGS, EPA-EMAP, RCRA, CERCLA, CWA, etc.)?
- Is precision being controlled in the context of repeatability, reproducibility, reproducibility plus, or some combination (specify)?
- How will precision be calculated and reported?
- Will the raw number results be reported in addition to summary statistics like RPDs, RSDs, sample standard deviations, or multiples of standard deviations? The raw numbers should also be reported so that in the future others can look at the values from different angles.
- How often will precision be estimated and reported?
- If the MQO is not met, what data will be rejected (often all data back to the last time the MQO was met, sometimes all data associated with that batch, that trip, or that day, whatever is logically associated with the QC sample that did not meet MQOs).

Standard precision measurement quality objectives (MQOs) can often be summarized in a table, sometimes along with systematic error, method detection limits, and blank control MQOs (see SFAN freshwater quality protocol QA/QC SOP example,

for example, see Table 9, on the Internet at http://www1.nature.nps.gov/im/units/sfan/reports/WQ/SOP4_QAPP_V2_01.pdf).

Precision in context of repeatability is the scenario where nothing in the measurement process changes. For chemical lab measurements, repeatability MQOs are typically used, but if multiple labs, multiple instruments, or multiple staff become factors, it becomes precision in the context of reproducibility.

Precision in the context of reproducibility is often very relevant for long term monitoring, since there will typically be changes in staff and instruments. Sometimes different staff and different instruments are even used by the same network during one season. In all cases where something in the measurement process changes, precision is controlled in the context of reproducibility. Changes should also trigger efforts to see if the changes introduces measurement bias (see section below on archiving bias from changes in the data analysis SOP).

"Precision Reproducibility Plus" is our NPS terminology for field duplicates when two samples that are not exactly the same are taken in close proximity in time and/or space. Since the samples may not be identical, the "plus" part of the phrase is a tip off that an additional potential source of variability is present. In this case, two potential sources of variability are being lumped, lack of prefect measurement precision plus potential true sample heterogeneity. Another way to say this is the precision plus (simply called field duplicates by some) covers both the precision of the sampling process and analytical precision. This is an acceptable approach, but planners should keep in mind that it may eventually trigger additional work compared to the more conventional estimates of measurement precision as repeatability, typically taken from measuring one homogenous sample twice. We are not suggesting that two different kinds of precision samples should be taken. However, if precision plus is the only kind done, and precision measurement quality objective goals start to be missed, it should trigger another step. In that case, the only way to determine how much of the variability is due to random error in the measurement process vs. how much is due to true sample heterogeneity would be to perform some repeatability precision checks (where only one sample is measured and nothing in the measurement process changes). To avoid having to do two types of precision, just doing precision in the context of repeatability is one acceptable approach.

Regardless of the type of precision controlled, precision QC samples are usually performed as duplicate samples every 10-20 samples (or every sampling batch, or every field sampling day, or each laboratory batch). Specify the details.

The text should document the extent to which precision MQOs will be used as data rejection criteria. Unless otherwise justified, they should usually be data rejection criteria. Here is an example. Suppose the precision MQO for a particular parameter is that a relative percent difference (RPD) of two duplicates cannot exceed a 30%. If the RPD exceeds that value, all values associated with that batch (or that QC sample) should be discarded. Recalibration or other adjustments should then be done until the MQO can be met.

Many types of biological inventory and monitoring projects have not historically estimated measurement precision. However, there is growing awareness of the need to do so. One can usually find a common-sense way to control and document measurement precision in biological projects. Often one can simply measure something twice to get a duplicate answer (see Part B for more detail).

As with other QA/QC topics, the word "precision" is sometimes used wrongly. It has too often been used for concepts other than variability. Some have used the word for the size of a confidence interval or for accuracy. Such usages of the word precision should be discouraged to prevent confusion. In concert with NIST/ISO definitions, precision is about variability (scatter), and is clearly not synonymous with confidence or accuracy.

Measurement Systematic Error/Bias/Percent Recovery (Still Wrongly Called "Accuracy" by Some)

Systematic error/bias is the systematic or persistent distortion of a measurement process that causes errors in only one direction (usually high or low). On the measurement scale of concern, systematic error and bias are usually considered synonyms. Though it has commonly been used in this context in the past, the word accuracy should not be used for the concept of bias. Uncertainty in accuracy can only be estimated after factoring in not only systematic error/bias, but also precision.

In water quality and contaminants work, systematic error/bias is usually expressed as a as % recovery (with the correct or expected answer being considered 100%) of an interval (such as 80-120%). That particular example could be shortened to 20%, but only if the QA/QC SOP makes it clear that 20% means a percent recovery interval of $\pm$ 20% (same as 80-120% recovery). The raw values used to calculate these percentages should also be reported to allow one to later look at the results in other ways (such as long term or multi-lab means or standard deviations).

For each measurement done in the field or lab, are the following adequately covered?

- A systematic error/bias measurement quality objective (MQO), such as % recovery must be within 80-120%.
- Will the MQO be used as a data acceptance performance standard?
- What is the data comparability source of the MQO (state, USGS, EPA-EMAP, RCRA, CERCLA, CWA, etc.)?
- How will systematic error/bias be calculated and reported?
- How often will systematic error/bias be estimated and reported?

Unless otherwise justified, MQOs should be data rejection criteria. So if the MQO for a particular parameter is that a % recovery cannot be worse that 70-130%, then if the recovery is 60% or 140%, all values associated with that batch should be discarded rather than reported into a long term database. In that case, recalibration or other adjustments in the measurement process should be made until the MQO can be met.

If one value (say water color by remote sensing) is being measured to estimate another value (say chlorophyll a, algae blooms, organic compounds like tannins, or mining wastes), how will bias and accuracy (including a precision component) and sensitivity be controlled and estimated? Will average observed to expected ratios be used? If so, how will they be used? Will root mean square error techniques be used? How? How will the sensitivity result compare to standard detection limits that use multiples of the standard deviation?

Many types of biological inventory and monitoring projects have not historically estimated measurement systematic error/bias. However, there is growing recognition of the need to do so, and one can usually find a common-sense way to control and document measurement bias in biological projects. One strategy sometimes used is to consider a senior expert's answer right or expected (100%) and a rookie trainee's answer wrong (see Part B for recommended strategies).

In forestry, different terms are used, but bias is controlled. Sometimes, the mean observation minus the known true value is used to estimate bias. Bias in regression analyses can be especially problematic, and rumors indicate outliers are often just discarded (Chapter 5, Measures and Estimates in the Statistical Methods for Adaptive Management Studies of the British Columbia Ministry of Forests Research).

Blank Control Bias (usually applicable to chemical lab work only)

Blank QC samples (samples known to be free of the analyte being measured) are tested to see if they have been contaminated with the analyte during collection, handling, and processing steps. Contamination results in a positive bias in the concentration. Unless otherwise justified, for lab chemical measurements of toxic chemicals, metals, pesticides, or nutrients, MQOs for blank control shall be listed in the QA/QC SOP. Unless otherwise justified, blank samples should not include any samples having measurable concentrations of the analyte of interest (measurable above the qualitative method detection limit such as an MDL).

For each chemical measurement done in the lab, are the following adequately covered?

1. A blank control measurement quality objective (MQO), if applicable.
2. What types of blanks will be controlled (trip blanks, lab blanks, etc.)
3. Will the MQO be used as a data acceptance performance standard?
4. Will data reported be adjusted by adding concentrations found in blanks? If not, how will blank control be accomplished (reduce contamination and re-run the samples?).
5. What is the data comparability source of the MQO (state, USGS, EPA-EMAP, RCRA, CERCLA, CWA, etc.)?
6. How will blank control be calculated and reported?
7. How often will blank control be estimated and reported?

Biological inventory and monitoring projects have not historically done blank control. However, if the scenario of wrongly assigning a number value when the true value is zero seems likely, it might be possible to develop a common-sense way to control bias from blanks.

Include Calibration Details

Instrument calibration details should be included either in the QC SOP or in a separate calibration SOP (checklist, op. cit.). If these details are somewhere else (perhaps in the field or lab method SOPs), there should be a "point to" in the QC SOP so that the

reader will be able to find them. Often an "additional" calibration step should be made in the field to make sure instruments have not fallen out of calibration during transport.

OTHER SOPS RELATED TO QA/QC

Include a Data Analysis SOP

Are there recommendations for routine data summaries and statistical analysis to detect change? How often will reporting and trend analyses be done? Does this SOP or the protocol narrative describe the frequency of testing and review of protocol effectiveness?

The data analysis SOP should include a discussion of the data analyses. This should include: 1) statistics to be used, 2) who will do the analyses, 3) how often the analyses will be done, and 4) what is planned to ensure that adequate staff time and project funding is set aside for this very important task. Most of the proposed statistics should be worked out with a statistician before protocols & SOPs are completed.

Networks may wish to state that statistics will usually be handled according to recommendations (say which ones) of the following water quality text books:

> Helsel, D.R. and R.M. Hirsch. 2002. Statistical Methods in Water Resources. US Geological Survey Techniques of Water Resources Investigations, Book 4, Chapter A3. This textbook is available for free on the Internet at http://water.usgs.gov/pubs/twri/twri4a3/pdf/twri4a3.pdf.
>
> D. Helsel. 2005. Nondetects and Data Analysis: Statistics for Censored Environmental Data. Wiley (op. cit., see detection limit section).
>
> McBride, G.B. 2005. Using Statistical Methods for Water Quality Management: Issues, Options and Solutions. Wiley, NY, 313 pp., http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471470163.html.
>
> EPA explains many probabilistic analysis issues at an EMAP monitoring design and analysis home page (http://www.epa.gov/nheerl/arm/designpages/design&analysis.htm). This is a good reference but probably cannot be used as a stand alone (don't just say analyzed according to EMAP suggestions).

Include a Cumulative Measurement Bias SOP

Is our internal NPS data from both old and newer measuring systems comparable enough that the different sets of data could be combined for purposes of determining trends or making management or regulatory decisions?

"Do not be swayed by the argument that we cannot change now because 'we have 4 years of data that will be compromised.' You are in this for the long haul. If you do not correct mistakes now, 25 years latter you will (should) be cursed" (Quote from Ken

Burnham at 2004 Denver LT Monitoring Conference, see: http://www.stat.colostate.edu/~nsu/starmap/pps/burnham.msts.pdf).

Often due to inevitable changes in staff and measuring equipment in long term monitoring, the data are not comparable enough to differentiate trends from changes caused by changes in measurements instruments, staff, or personnel. Therefore, overlapping measures need to be done for a period of time to see if measurement bias has been introduced from the old measurements to the new.

WRD as well as VS/NRPC Database Staff (Margaret Beer) believe this information is important enough to warrant its own SOP. It would also be acceptable for the monitoring network to choose to document this type of information in the Data Analysis SOP, if a good rationale for doing so is provided. However, for NPS VS consistency, we recommend that a separate Cumulative Measurement Bias SOP be included.

Either way, the information is important enough for those who will eventually be trying to detect trends that liberal use of "point-too links" should be included in the Data Analysis SOP, so that future data users can find this information.

Method, equipment, and personnel changes are inevitable in long term monitoring. The requirement of overlapping old and new measurement methods is in Oakley et al. (op. cit.). However, both this requirement and the underlying reason for it are too often overlooked.

The SOP should detail how long the old and new methods are to be overlapped to determine changes in measurement precision, measurement sensitivity, and (especially) measurement bias. The following text (or something like it and defendable) is suggested for inclusion in a Cumulative Measurement Bias SOP, with crosslink text in the data analysis SOP pointing to the SOP where this may be found:

When the Only Change is a Change in Personnel
Doing the Measurements, Observations, or Ratings:

Single (identical) samples will be measured by old and new personnel at least 7 times when the only thing changing is staff doing the measuring or observations.

When the Change is a Change in Meters,
Measurement Instruments, Methods, or SOPs:

At least 30 overlap measurements will be made when a method, SOP, meter, or measuring instrument changes.

When the Change is a Change in an Indicator
Or in One Surrogate Measure to Estimate Another

At least 50 overlap measurements will be made and results recorded. The bigger the method or SOP change, the more repeat sampling may be appropriate. Some states have gone to great lengths when replacing one indicator with another. For example, Oregon created regression derived equations for estimating fecal coliform values from *Escherichia coli* values (or vice versa) after converting to *E. coli* monitoring. They monitored both fecal and E. coli side by side for six years before

deciding that *E. coli* = about half fecal coliforms for Oregon's rivers and streams and becoming comfortable with dropping fecal coliforms (C.G. Cude. 2005. Accommodating change of bacterial indicators in long term water quality datasets. Jour. Am. Water Resources Assn. 41(1): 47-54). An indicator change is a bigger change than a staff change or a method change.

In all of the cases listed above, the following shall be archived in the Cumulative Measurement Bias SOP:

1. Average % bias change from the old measurement system to the new, calculated as an average of percent differences (PDs, not relative percent differences or RPDs)
2. There should be a clear statement of which way the bias went. If the values for the new measurement system are on average higher than those for the old (PDs are on average positive), the bias resulting from the change is positive. The PD change is calculated by subtracting the old measure from the new one, then dividing the difference times the old measure, then taking the result times 100. By subtracting the old measure from the new one, if the change is positive, the PD calculated value will be positive too. Trends are then based on values normalized to the original numbers. Fraction of change from old to new is calculated as old value divided by new. New values can then be normalized to old by multiplying the new values times the calculated fraction of change.
3. A 95% t-distribution confidence interval about the mean for the average bias change expressed as a percent difference (PD)
4. The sample standard deviation used to estimate the average
5. The sample size used to estimate the above (the number of overlapping measures used to calculate the average PD),
6. The precision as reproducibility expanded uncertainty for both the old and new measurement systems (see separate explanations herein)
7. Measurement sensitivity, as either an MDL (if some measurements are near or below the MDL) or AMS (if all measurements are well above the MDL or if the MDL is otherwise not appropriate, see separate discussions herein) for both the old and new measurement systems
8. The date that the overlapping measurements started
9. The date that the overlapping measurements stopped
10. The date that the average PD bias change from old to new measurements was calculated.
11. All paired raw values, should future statistician desire to normalize values a different way when estimating trends.

In all of the cases listed above, if the 95% t-distribution confidence interval about each side of the mean for the average bias change (expressed as a percent difference) exceeds 20% of the mean of the either the old or new measurements (not 20% of the average PD bias change), the number of overlap measures will be increased until the confidence interval is no greater than 20% or less of the mean. The SOP should also state

that if this cannot be accomplished, changes in the measurement process (better calibration, better training, etc.) will be made until it can be accomplished.

The "extra" cost of the overlap samples is a practical issue. The 20% decision rule can also be used to decrease the number of overlap samples (from the 30-50 defaults recommended above), should the network have reason to believe that bias will be a small issue and that that cost of the overlap measures will be an major issue for any given measurement system change.

Why use 20% of the mean of either the first or last batch (whichever is greater) in the decision rule rather than 20% of the mean of all percent differences in the overlap measures comparison? The short answer is 1) because it has a better chance of environmental relevance from the standpoint of how big of a change do we really care about and generic/typical signal to noise ratio rationales, 2) because it requires lower sample sizes (and lower costs) for the overlap (paired) measures, and 3) because it doesn't matter that the average bias be extremely well estimated, just that it be well quantified and that the quantification estimate is controlled within logical and stated limits. In other words, using the decision rule above, the bias estimates are not uncontrolled. This is an improvement over the past typical practices and the required number of overlap measurements is not excessive. If there are logical reasons why the estimate should be controlled even more tightly (for example, the measures are near a magnitude that would bring one is near a collapse threshold value for a rare resource), there is nothing to prevent a monitoring network from controlling the estimate more tightly. One logical way to do that would make the comparison to 20% of the mean of all percent differences in the comparison rather than 20% of the mean of either the original or old measurements in the overlap measures. Statisticians could also suggest other more stringent ways if asked. Controlling it more tightly would usually simply require larger sample sizes (sometimes much larger if the variability was high) for the number of overlap measures.

If the variability in the overlap samples is low (say for example a typical overlap comparison pair is 5.32 and 5.31), the required sample sizes for a more stringent decision rule would not be too high, if higher at all. On the other hand, if the parameter is something like an estimate of % cover of vegetation covering a stream, and different observers came up with very different estimates (for example 30% vs. 50%), it would be harder to control the bias estimates very tightly without a very large number of overlap estimates.

One side benefit of doing overlapping measurements is that one might discover the old meter or method is better and decide not to use the new one. Also, if a change in personnel changes the bias or precision in unacceptable ways, it is best to find that out as early as possible so that additional training or other changes can be made until an acceptable result is obtained.

Is the above too much to ask? We do not believe so, and the cost of not documenting cumulative measurement change bias is an inability to differentiate true environmental trends from measurement or estimation changes. We have seen dramatic examples where one could not differentiate trends from method changes in past data from the last 50 years, and we would like to avoid that with our new monitoring program. Why bother to monitor long term if we don't do it in ways that allow us to document true trends in defendable ways?

Even volunteer groups are performing these kinds of method change comparisons now. One expert recommended that volunteer groups overlap 30-50 paired observations when changing salinity methods (P. Bergstrom. 2005. Comparing Four Salinity Methods. 2005. The Volunteer Monitor 17 (1):21, http://www.epa.gov/volunteer/newsletter/volmon17no1.pdf). The NPS typically does not want to be less rigorous than volunteer groups; 25-50 is often a minimum sample size to estimate many summary statistics (such as proportions and means) well, so unless otherwise justified, overlap at least 30 samples for method or other substantial changes.

Even small changes in measurement bias can accumulate and become significant over time. "Point-to" notes about where such documentation is should also go with the data, as part of metadata notes or introductory notes.

The goal would be for someone 100 years later to be able to discover the effect of the various changes in measurement bias. These are the types of examples that a future data user would need to discover to enable that user to separate true trends from method changes: 1) 90 years ago there was a method change that resulted +2% change (on identical samples) from the previous method, 2) 80 years ago there was another method change that resulted in another change of +4% from the method used in the previous 20 years, 3) 60 years ago there was another method change that resulted in a new plus 3% bias, and 4) 75 years ago there was another method change that resulted in a -1% bias from the years just before. In this pretend example, unless the future data user could find this type of information, that person might conclude there was a steady upward trend that leveled off a bit at the end of the period, when the only changes were really a series of bias changes caused by method or SOP changes. This issue is important enough for long term monitoring that some redundancy provided by the multiple "point to" links from other places seems prudent.

How the data will be normalized could be handled in either SOP with "point-to" links from the other. Usually all data will be normalized to the original measurement method. For example, in 2100, data from 2006, 2020, and 2040, etc. might all be normalized to 2006 data equivalents. If it is determined that the original method used in 2006 was too deficient to form a defendable normalizing starting point for example, measurement uncertainty, measurement sensitivity, and/or measurement bias were so bad that (confidence intervals on the estimates were above 20% of the mean for example) or incompletely documented, then start over with a new normalization point (say 2020 or 2040) when the deficiencies in documentation of measurement performance have been corrected.

In summary, it is suggested that the cumulative results of the bias over the years be detailed in the Cumulative Measurement Bias SOP in each protocol with "point to" hyperlinks from other places people might look, such as the protocol revision log, each field and lab SOP for methods, the data management SOP, the data management section of the protocol narratives and central monitoring plan, the data acquisition parts of the central monitoring plan, the Data Analysis SOP, and the precision and bias discussions in the QC SOP.

Include STORET Details in a Data Management SOP

Documentation and planning in the SOP needs to include matching the network's characteristics/parameters with the official standardized EPA list of 337,378 (as of 1/5/2005) characteristics (found in tblDef_TSRCHAR in NPSTORET. For questions, contact Dean_Tucker@NPS.GOV) and or see http://nrdata.nps.gov/Programs/Water/storetcharacteristics/storetcharacteristics.zip). How the data collected will be archived in STORET and NPSTORET should be detailed in a Data Management SOP.

End of Part B lite. More detail on each of the topics in Part B lite is found in the long version of Part B at http://science.nature.nps.gov/im/monitor/protocols/wqPartB.doc.